

Achilles Thoma
Sheila Sprague
Sophocles H. Voineskos
Pablo E. Serrano
Charles H. Goldsmith
Editors

Evidence-Based Surgery

A Guide to Understanding and
Interpreting the Surgical Literature

Second Edition

Evidence-Based Surgery

Achilles Thoma • Sheila Sprague
Sophocles H. Voineskos • Pablo E. Serrano
Charles H. Goldsmith
Editors

Evidence-Based Surgery

A Guide to Understanding
and Interpreting the Surgical
Literature

Second Edition

 Springer

Editors

Achilles Thoma
Division of Plastic Surgery, Department
of Surgery & Department of Health
Research Methods, Evidence and Impact
McMaster University
Hamilton, ON, Canada

Sophocles H. Voineskos
Division of Plastic, Reconstructive, and
Aesthetic Surgery, Department of
Surgery
University of Toronto, & Li Ka Shing
Knowledge Institute, Unity Health
Toronto, ON, Canada

Charles H. Goldsmith
Department of Health Research
Methods, Evidence and Impact
McMaster University
Hamilton, ON, Canada

Sheila Sprague
Division of Orthopaedic Surgery,
Department of Surgery & Department of
Health Research Methods, Evidence and
Impact, McMaster University
Hamilton, ON, Canada

Pablo E. Serrano
Division of General Surgery,
Department of Surgery & Department of
Health Research Methods, Evidence &
Impact, McMaster University
Hamilton, ON, Canada

ISBN 978-3-031-87082-8 ISBN 978-3-031-87083-5 (eBook)
<https://doi.org/10.1007/978-3-031-87083-5>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2019, 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

We dedicate this book to all learners, surgical residents, and surgeons who have adopted evidence-based surgery principles, thus enhancing surgical care in their institutions and communities.

Foreword

The first edition of *Evidence-Based Surgery* was an excellent resource that encapsulated fundamental principles and the philosophy around understanding and interpreting the surgical literature. It emphasized the critical importance of evidence in guiding surgical decision making and the need to adapt the principles of EBM into all areas of surgery including clinical care, research, and education. The second edition builds on this foundation to create the go-to resource for those in surgery with an interest in being well-informed regarding changes in various approaches to evidence creation or being at the forefront of grasping advances in numerous new and older study designs. An incredibly wide and comprehensive range of information is presented that is of critical importance but easy to read and understand that leaves the reader wanting to finish the textbook in a single sitting.

With the explosive growth in surgical evidence and surgical trials, the second edition of *Evidence-Based Surgery* is much needed and has been highly anticipated. Each chapter in this textbook has been enhanced and standardized to improve its readability with a glossary of terms, tables, figures, and an appendix which emphasize key points presented in the chapter. Moreover, the chapters are updated with new content that makes the book fresh with a strong focus on key issues, controversies, and new approaches. The writing style will resonate with learners at all levels including students, residents, fellows, research personnel, and surgeons, using a how-to approach that will enable the effective practice of evidence-based surgery as it applies to clinical care and research. Today's learner not only wants to understand the evidence but also the methodology associated with its creation.

The textbook contains 34 chapters and provides a comprehensive overview on key areas such as trial design, methodology, interpretation of the literature, outcomes, and statistics. The editors and authors are well-known and highly regarded experts in the field and cover critically important issues with direct impact on the practice of surgery including the evaluation of surgical interventions, quality improvement and patient safety and harm, economic evaluations, guidelines, and enhanced recovery after surgery. The chapters on randomized controlled trials are outstanding and provide an essential resource for understanding the depth and breadth of options in this area. Finally, there are four new chapters that highlight key new areas of interest including Cluster RCTs, Network Meta-analysis, and Core Outcome Sets. Without any gaps, the editors are to be congratulated for their significant contribution as

this textbook will bring the reader back over and over with its ability to serve as a comprehensive source of highly relevant information.

Richard Ivey Professor and Chair
Distinguished University Professor
Western University
London, ON, Canada

Emil H. Schemitsch, MD FRCS(C)

Preface

In our view the publication of the second edition of *Evidence-Based Surgery: Understanding and Interpreting the Surgical Literature* is of great relevance to surgeons of all specialties.

Although the terms Evidence-Based Medicine (EBM) and Evidence-Based Surgery (EBS) are mentioned often in the literature, there persists a gap in the understanding of what these terms and their underlying philosophy mean despite the introduction of this new method of learning and teaching three decades earlier.

The first and second edition have their roots in the EBS surgery workshops we undertook to teach surgeons at McMaster University and other institutions since the 1990s.

Since the first edition was published, there have been methodological advances in the various study designs. Large surgical trials that use new ways of data acquisition or that demonstrate different methods of obtaining informed consent have been developed. New types of surgical clinical trials have been developed, especially in surgical oncology, involving multidisciplinary teams (i.e., radiation oncology or medical oncology). These will be addressed in the individual chapters.

We have added four new chapters in the 2nd edition. The chapter on Cluster RCTs will complement the other Randomized Controlled Trial (RCT) designs (superiority RCT, non-inferiority RCT, expertise-based RCTs, and Pilot RCT). The chapter on Network Meta-analysis will complement the chapter of Systematic Reviews and meta-analysis. The chapter of Enhanced Recovery After Surgery (ERAS) is an important advance in standardizing surgical care in many specialties and hence of clinical relevance. We have added a chapter on Core Outcome Sets (COS). As the inclusion of critically important outcomes in any surgical research project is of paramount importance for credible research, we believe that surgeons should be familiar with this presently evolving area.

In the second edition we made some aesthetic changes to improve the readability of the figures. Each chapter will start with a glossary of the terms used in the chapter and will include as an appendix the literature search strategy for interested readers who may wish to replicate the findings.

Another big change from the first edition is the addition of Pablo E. Serrano as a fifth co-editor who provides the perspective of a new generation of surgeon methodologists with emphasis in surgical oncology, bringing the basic concepts of surgical trials to oncology research.

We hope that you will enjoy the second edition, and we welcome feedback on this work.

Achilles Thoma
Sheila Sprague
Sophocles H. Voineskos
Pablo E. Serrano
Charles H. Goldsmith

Acknowledgments

This book would not have been made possible without the administrative and editorial assistance of Emily Dunn (McMaster University), and Janakiraman Ganesan (Springer). We thank them for their key contributions in bringing this work to fruition.

Contents

1 History of Evidence-Based Surgery (EBS)	1
Achilles Thoma, Jessica Murphy, Emily Dunn, Sheila Sprague, and Charles H. Goldsmith	
2 The Steps of Practicing Evidence-Based Surgery (EBS).	11
Achilles Thoma, Emily Dunn, Sheila Sprague, Luis H. Braga, and Sophocles H. Voineskos	
3 Developing a Surgical Clinical Research Question: To Find the Answer in Literature Search or in Pursuing Clinical Research	21
Achilles Thoma, Emily Dunn, Sheila Sprague, Sophocles H. Voineskos, and Jessica Murphy	
4 Finding the Evidence Through Searching the Literature	27
Laura Banfield and Jo-Anne Petropoulos	
5 Hierarchy of Evidence in Surgical Research	43
Gina Del Fabbro, Sofia Bzovsky, Achilles Thoma, and Sheila Sprague	
6 Evaluating Surgical Interventions.	55
Tyler McKechnie, Yung Lee, Aristithes G. Doumouras, and Dennis Hong	
7 A Primer on Outcome Measures	69
Joy MacDermid	
8 Patient-Reported Outcome Measures in Surgery: Development, Selection, and Evaluation.	75
Manraj N. Kaur, Lucas Gallo, Anne F. Klassen, and Andrea L. Pusic	
9 Surrogate Endpoints.	95
Dan Cohen, Japleen Thind, Nicole Simunovic, and Olufemi Ayeni	
10 How to Assess an Article that Deals with Health-Related Quality of Life	103
Achilles Thoma, Jenny Santos, Margherita Cadeddu, Eric Duku, Emily Dunn, and Charles H. Goldsmith	

11	Randomized Controlled Trial Comparing Surgical Interventions	113
	Lily J. Park, Marko Simunovic, Mohit Bhandari, and Pablo E. Serrano	
12	Users' Guide to the Surgical Literature: How to Assess a Pilot Trial in Surgery	125
	Guowei Li, Xuerui Bai, Gillian A. Lancaster, and Lehana Thabane	
13	Assessing Noninferiority Randomized Controlled Surgical Trials	139
	Alla Bengizi, Mansi Patel, Sandra Ofori, Yaadwinder Shergill, and Jason W. Busse	
14	Expertise-Based Randomized Controlled Trials	151
	Tyler McKechnie, Daniel Waltho, Kristen Davidge, and Cagla Eskicioglu	
15	Cluster-Randomized Trials in Surgery	163
	Kyle Gouveia, Sofia Bzovsky, and Sheila Sprague	
16	The Surgeon's Guide to Systematic Review and Meta-Analysis	177
	Lucas Gallo, Patrick Kim, Andrea Copeland, Noor Alolabi, and Achilles Thoma	
17	Prospective and Retrospective Cohort Studies	191
	Richard Walker, Ramy Behman, Lev Bubis, and Paul Karanickolas	
18	Case-Control Studies	205
	Achilles Thoma, Jenny Santos, Jessica Murphy, Eric Duku, Emily Dunn, and Charles H. Goldsmith	
19	Evaluating Case Series in Surgery	217
	Christopher J. Coroneos and Brian Chin	
20	Quality Improvement and Patient Safety in Surgery	227
	Jessica H. Hannick and Martin A. Koyle	
21	Diagnostic Studies in Surgery	235
	Michael Xie, Stuart Archibald, Achilles Thoma, Michael K. Gupta, Jessica Murphy, and Charles H. Goldsmith	
22	How to Assess a Prognostic Study	255
	Saurabh Gupta, Kevin Kim, Emilie Belley-Côté, and Richard P. Whitlock	
23	Decision Analysis and Surgery	263
	Alison B. Suraci and Gloria M. Rockwell	

24	Economic Evaluations in Surgery	277
	Achilles Thoma, Feng Xie, Emily Dunn, Jenny Santos, and Charles H. Goldsmith	
25	Studies Reporting Harm in Surgery	291
	Lily J. Park, Marko Simunovic, Pablo E. Serrano, Mohit Bhandari, and Robin McLeod	
26	Evaluating Surveys and Questionnaires in Surgical Research.	301
	Brian Chin and Christopher J. Coroneos	
27	Simple Statistical Tests and <i>P</i> Values.	313
	Charles H. Goldsmith, Eric Duku, Achilles Thoma, Jessica Murphy, and Emily Dunn	
28	Confidence Intervals.	329
	Lily J. Park, Jessica Bogach, Lawrence Mbuagbaw, and Margherita Cadeddu	
29	Power and Sample Size.	341
	Eric Duku, Jessica Murphy, Emily Dunn, Achilles Thoma, and Charles H. Goldsmith	
30	Subgroup Analyses in Surgery.	357
	Chloe R. Wong, Pablo E. Serrano, Alexandra C. Hatchell, and Sophocles H. Voineskos	
31	Surgical Guideline Appraisal.	371
	Bright Huo, Christopher J. Coroneos, Neil Smart, Cesare Stabilini, Melissa Brouwers, and Stavros A. Antoniou	
32	Network Meta-analyses in Surgical Trials	383
	Lily J. Park, Kevin Um, Dena Zera, Emmanuelle Duceppe, and Pablo E. Serrano	
33	Enhanced Recovery After Surgery (ERAS)	397
	Kathryn J. Uhlman, Alexandra C. Hatchell, Achilles Thoma, and Claire F. Temple-Oberle	
34	Appraising Core Outcome Sets (COS) in Surgical Research . . .	407
	George Elzawy, Achilles Thoma, Yaeesh Sardiwalla, Jessica Murphy, Marta Karpinski, and Richard P. Whitlock	
	Index	415

Contributors

Noor Alolabi Division of Plastic and Reconstructive Surgery, Oakville
Trafalgar Memorial Hospital, Oakville, ON, Canada

Faculty of Health Sciences McMaster University, Hamilton, ON, Canada

Stavros A. Antoniou Department of General Surgery, Papageorgiou General
Hospital, Thessaloniki, Greece

EAES Guidelines Committee, Eindhoven, Netherlands

Stuart Archibald Division of Head and Neck Surgery, Department of
Surgery, McMaster University, Hamilton, ON, Canada

Olufemi Ayeni Division of Orthopaedic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada

Xuerui Bai Department of Epidemiology, School of Medicine, Jinan
University, Guangzhou, China

Laura Banfield Health Science Library, HSC 2B, McMaster University,
Hamilton, ON, Canada

Ramy Behman Division of Surgery, University of Texas MD Anderson
Cancer Center, Houston, TX, USA

Emilie Belley-Côté Department of Health Research Methods, Evidence,
and Impact, McMaster University, Hamilton, ON, Canada

Population Health Research Institute, Hamilton, ON, Canada

Division of Cardiology & Critical Care, Department of Medicine, McMaster
University, Hamilton, ON, Canada

Alla Bengizi Department of Health Research Methods, Evidence, and
Impact, McMaster University, Hamilton, ON, Canada

Mohit Bhandari Department of Surgery, Division of Orthopaedic Surgery,
McMaster University, Hamilton, ON, Canada

Jessica Bogach Division of General Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada

Department of Health Research Methods, Evidence, and Impact, McMaster
University, Hamilton, ON, Canada

Luis H. Braga Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Division of Urology, Department of Surgery, McMaster University, Hamilton, ON, Canada

Melissa Brouwers School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

Lev Bubis Department of Surgery, University of Manitoba, Health Sciences Centre, Winnipeg, MB, Canada

Jason W. Busse Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Department of Anesthesia, McMaster University, Hamilton, ON, Canada

Sofia Bzovsky Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Margherita Cadeddu Division of General Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Brian Chin Division of Plastic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Dan Cohen Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Andrea Copeland Breast Reconstruction and Microsurgery Fellow, Royal Brisbane and Women's Hospital, Brisbane, QLD, Australia

Christopher J. Coroneos Division of Plastic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Kristen Davidge Division of Plastic Surgery, Department of Surgery, University of Toronto, Toronto, ON, Canada

Gina Del Fabbro Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Aristithes G. Doumouras Department of Surgery, McMaster University, Hamilton, ON, Canada

Department of Surgery, St. Joseph's Healthcare, Hamilton, ON, Canada

Emmanuelle Duceppe Department of Medicine, Department of Internal Medicine, CHUM, Montreal, ON, Canada

Eric Duku Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Offord Centre for Child Studies, Department of Psychiatry and Behavioral Neuroscience, McMaster University, Hamilton, ON, Canada

Emily Dunn Division of Plastic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

George Elzawy Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

Cagla Eskicioglu Division of General Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Lucas Gallo Division of Plastic and Reconstructive Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Charles H. Goldsmith Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada

Kyle Gouveia Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Michael K. Gupta Division of Head and Neck Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Saurabh Gupta Department of Surgery, McMaster University, Hamilton, ON, Canada

Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Jessica H. Hannick Department of Urology, Cleveland Clinic Children's Hospital, Cleveland, OH, USA

Alexandra C. Hatchell Departments of Surgery and Oncology, Section of Plastic Surgery, University of Calgary, Calgary, AB, Canada

Department of Surgery, University of Calgary, Calgary, AB, Canada

Department of Oncology, University of Calgary, Calgary, AB, Canada

Dennis Hong Department of Surgery, McMaster University, Hamilton, ON, Canada

Department of Surgery, St. Joseph's Healthcare, Hamilton, ON, Canada

Bright Huo Division of General Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Paul Karanicolas Department of Surgery, University of Toronto, Sunnybrook Health Sciences Centre, Sunnybrook Research Institute, Toronto, ON, Canada

Marta Karpinski Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

Manraj N. Kaur Patient-Reported Outcomes, Value & Experience (PROVE) Center, Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Kevin Kim Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Population Health Research Institute, Hamilton, ON, Canada

Patrick Kim Division of Plastic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Anne F. Klassen Department of Pediatrics, McMaster University, Hamilton, ON, Canada

Martin A. Koyle Professor Emeritus, Temerty Faculty of Medicine and Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, ON, Canada

Gillian A. Lancaster School of Medicine, Keele University, Keele, UK

Yung Lee Department of Surgery, McMaster University, Hamilton, ON, Canada

Guowei Li Center for Clinical Epidemiology and Methodology (CCEM), The Affiliated Guangdong Second Provincial General Hospital of Jinan University, Guangzhou, China

St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada

Joy MacDermid Roth | McFarlane Hand and Upper Limb Centre, St Joseph's Health Centre, London, ON, Canada

School of Physical Therapy and Department of Surgery, Western University, London, ON, Canada

Lawrence Mbuagbaw Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Tyler McKechnie Division of General Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Robin McLeod Department of Surgery and the Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

Jessica Murphy Division of Physical Medicine & Rehabilitation, Department of Medicine, McMaster University, Hamilton, ON, Canada

Sandra Ofori Department of Medicine, McMaster University, Hamilton, ON, Canada

Lily J. Park Division of General Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Mansi Patel Department of Surgery, McMaster University, Hamilton, ON, Canada

Jo-Anne Petropoulos Health Science Library, HSC 2B, McMaster University, Hamilton, ON, Canada

Andrea L. Pusic Patient-Reported Outcomes, Value & Experience (PROVE) Center, Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Gloria M. Rockwell Department of Surgery, Division of Plastic Surgery, University of Ottawa, Ottawa, ON, Canada

Jenny Santos Southwestern Public Health (SWPH), St. Thomas, ON, Canada

Yaesh Sardiwalla Division of Plastic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Pablo E. Serrano Division of General Surgery, Department of Surgery & Department of Health Research Methods, Evidence & Impact, McMaster University, Hamilton, ON, Canada

Yaadwinder Shergill Ottawa Hospital Research Institute, Ottawa, ON, Canada

Marko Simunovic Division of General Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Nicole Simunovic Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Neil Smart Department of Gastrointestinal Surgery, Royal Devon & Exeter Hospital, Exeter, UK

Sheila Sprague Division of Orthopaedic Surgery, Department of Surgery & Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada

Cesare Stabilini Department of Surgical Sciences (DISC), University of Genoa, Genoa, Italy

Alison B. Suraci Department of Surgery, Division of Orthopaedic Surgery, Banner Medical Group, Phoenix, AZ, USA

Claire F. Temple-Oberle Department of Surgery, University of Calgary, Calgary, AB, Canada

Department of Oncology, University of Calgary, Calgary, AB, Canada

Lehana Thabane St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada
Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Japleen Thind Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

Achilles Thoma Division of Plastic Surgery, Department of Surgery & Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada

Kathryn J. Uhlman Division of Plastic, Reconstructive and Aesthetic Surgery, Department of Surgery, University of Toronto, Toronto, ON, Canada

Kevin Um Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Department of Medicine, Division of Cardiology, McMaster University, Hamilton, ON, Canada

Sophocles H. Voineskos Division of Plastic, Reconstructive, and Aesthetic Surgery, Department of Surgery, University of Toronto, & Li Ka Shing Knowledge Institute, Unity Health, Toronto, ON, Canada

Richard Walker Department of Surgery, University of Toronto, Toronto, ON, Canada

Daniel Waltho Department of Surgery, Division of Plastic Surgery, Oakville-Trafalgar Memorial Hospital, Oakville, ON, Canada

Richard P. Whitlock Division of Cardiac Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Population Health Research Institute, Hamilton, ON, Canada

Chloe R. Wong Department of Surgery, Division of Plastic, Reconstructive and Aesthetic Surgery, University of Toronto, Toronto, ON, Canada

Feng Xie Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Centre for Health Economics and Policy Analysis, McMaster University, Hamilton, ON, Canada

Michael Xie Division of Head and Neck Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Dena Zera Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

History of Evidence-Based Surgery (EBS)

1

Achilles Thoma, Jessica Murphy, Emily Dunn,
Sheila Sprague, and Charles H. Goldsmith

Abbreviations

EBS	Evidence-based surgery	WHO	World Health Organization
EBM	Evidence-based medicine	COSMIN	Consensus-based Standards for the selection of health Measurement INstruments
CMAJ	Canadian Medical Association Journal	EQUATOR	Enhancing the QUALity and Transparency Of health Research
JAMA	Journal of the American Medical Association	PRISMA	Preferred reporting items for systematic reviews and meta-analyses
LOE	Levels of evidence	COMET	Core outcome measures in effectiveness trials
DCRI	Duke Clinical Research Institute	FDA	Food and drug administration
SCCR	Stanford University Center for Clinical Research	NLM	National Library of Medicine
SRM	Surgical research methodology	NIH	National Institutes of Health
OMERACT	Outcome measurements in rheumatology	PROSPERO	International prospective registrar of systematic reviews
GRADE	Grading of recommendations assessment, development and evaluation		

A. Thoma (✉)
Division of Plastic Surgery, Department of Surgery
& Department of Health Research Methods, Evidence
and Impact, McMaster University, Hamilton, ON,
Canada
e-mail: athoma@mcmaster.ca

J. Murphy
Division of Physical Medicine & Rehabilitation,
Department of Medicine, McMaster University,
Hamilton, ON, Canada
e-mail: murphj11@mcmaster.ca

E. Dunn
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: dunne4@mcmaster.ca

S. Sprague
Division of Orthopaedic Surgery, Department of
Surgery & Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: sprags@mcmaster.ca

C. H. Goldsmith
Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: goldsmi@mcmaster.ca

Evidence-based surgery (EBS), which evolved from its precursor, evidence-based medicine (EBM), represents a paradigm shift in surgical practice. First coined by Gordon Guyatt in the early 1990s [1], EBM describes an approach to decision-making in medical practice that emphasizes the utilization, and critical appraisal, of the best available evidence for patient care. Before the development of EBM, physicians from the time of Hippocrates learned the craft of medicine through experience; the art and knowledge of which was passed from teacher to student. The EBM movement was introduced to address this gap, with a goal to ensure that patients were treated based on evidence, not the word of authorities [2].

While EBM, as a movement, is relatively new to the medical community, clinical trials have been taking place in some form since the 1500s. It has been reported that the first clinical trial was conducted, accidentally, by Ambroise Pare while treating military soldiers [3]. When the supply of oil, the standard treatment for wounds, was sparse he mixed a digestive of egg yolks, rose oil, and turpentine. Pare reported the differences between the two treatments, with those receiving the new treatment experiencing “little pain and decreased swelling” [3]. The first physician-conducted controlled clinical trial within the modern era was recorded 200 years later by James Lind. Lind planned a trial comparing cures for scurvy using 12 patients at sea [3]. From this study, it was discovered that citrus fruits could treat scurvy; however, it was not until 50 years later that the British Navy made lemon juice a part of sailors’ diets [3].

Since the work of James Lind, several notable events have occurred that have contributed to the development of science-based surgery and medicine, including:

1. The careful documentation of surgical interventions and results by Billroth and Halstead [4]
2. The publication of the Flexner report in the early twentieth century, cementing “scientific inquiry as the bedrock of American medicine” [5]
3. The US FDA Kefauver-Harris Act in the 1960s, demanding rigorous empirical testing

of clinical trials in humans leading to the established claims regarding drug efficacy and safety of new pharmaceutical innovations [5]

The “experiential” nature of the development of surgery in modern times was highlighted by Meakins [6]. He believed that surgical training was performed “in a hierarchical environment” where the professor or chief likely defined the way clinical situations were to be managed, and how the operation was to be done [6].

Sometime within the 1970s and 1980s, a paradigm shift occurred in which evidentiary rules changed the hierarchical system mentioned above by Meakins; this new paradigm placed a higher value on evidence than authority. The protagonists in this paradigm shift were a group of clinical epidemiologists, including David Sackett (considered by most, the father of EBM), David Eddy, Archie Cochrane, and others. They highlighted the need for strengthening the empirical evidence of medicine [5–11]. They proposed the initial evidentiary rules for guiding clinical decisions, and shortly after, published the first of a series of articles in the *Canadian Medical Association Journal* (CMAJ) advising physicians how to appraise the medical literature [12]. This series of articles was followed by the well-known article series, Users’ Guides to the Medical Literature, first published in the *Journal of the American Medical Association* (JAMA) in 1993 [13].

Some may argue that the seeds of the EBM movement started with the publication of a rudimentary form of levels of evidence by the Canadian Task Force on Periodic Health Examination in 1976 [7]. This Task Force proposed a rating system with four levels of evidence (LOE) and the corresponding types of evidence (see Fig. 1.1). This LOE rating system, which is central to the philosophy of EBM and EBS, was later improved by Sackett to include five levels (see Fig. 1.2) [8]. These early LOE rating scales have been since modified and made more stringent to consider a study’s methodological quality [14].

The advent of the EBM movement in the early 1980s was slowly adopted by surgery. This is, more than likely, because the originators of this EBM movement were primarily internists, and

Fig. 1.1 CTFPHE's examination's levels of evidence. (Adapted from the CTFPHE [7])

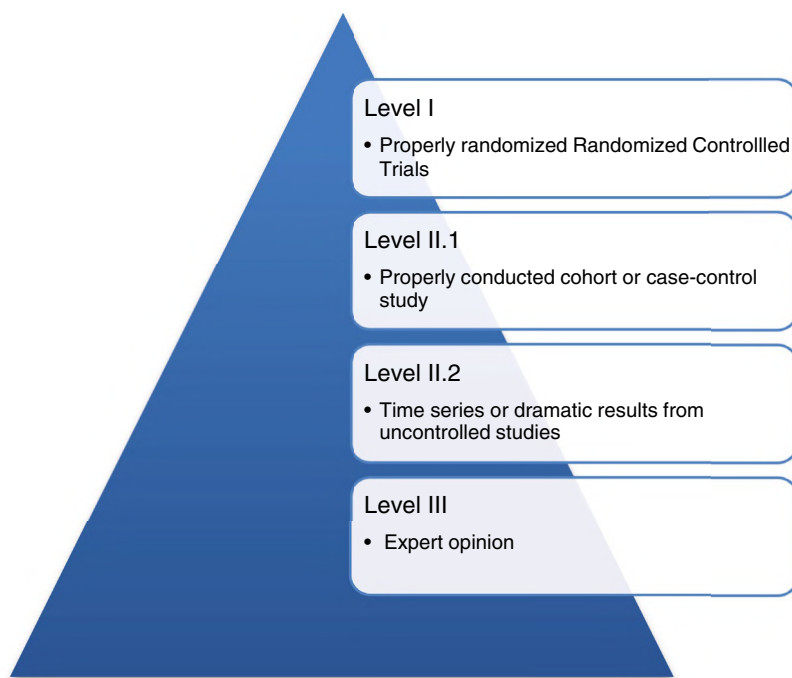
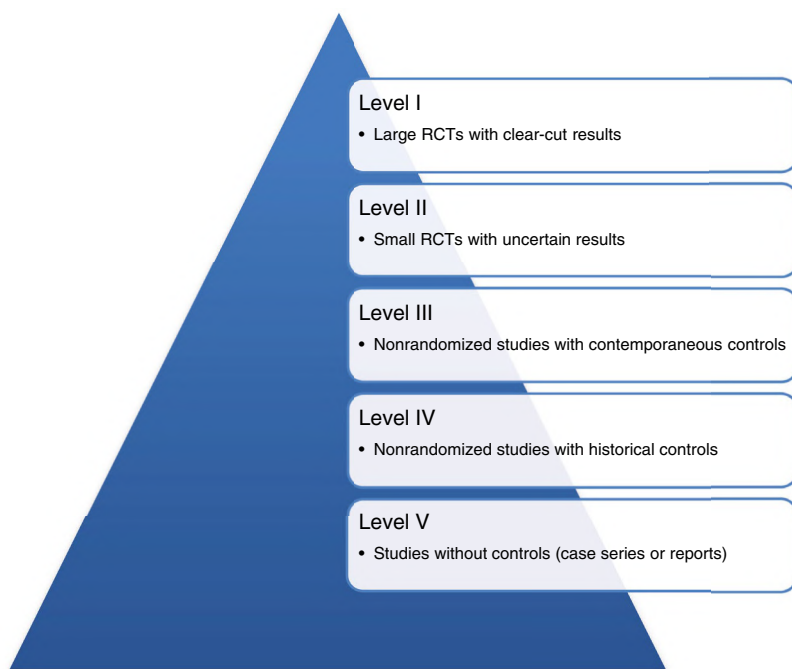


Fig. 1.2 David Sackett's levels of evidence. (Adapted from Sackett et al. [8])



their concepts focused on clinical scenarios in internal medicine. To introduce the principles of EBM into surgery, the definition of EBM was slightly altered to EBS by a group of surgeons and methodologists at McMaster University [15, 16]. Evidence-based surgery (EBS) is thus defined as the integration of best research evidence, patients' preferences/values, health resource availability, clinical setting, and ultimately our surgical expertise. This group of surgeons and methodologists, including mentees and colleagues of David Sackett published a series of articles in the *Canadian Journal of Surgery* with the goal to teach surgeons of all subspecialties the skills of appraising a surgical article before applying its conclusions to their practice [15–35]. Gradually, some surgical subspecialties encouraged their members to adopt the principles of EBS [36, 37]. There is an impression that the increase and quality of surgical research seen since 2000 in these subspecialties is due to the adoption of the EBM and EBS philosophies.

In practicing EBS, we base our decisions on the five pillars, as shown in Fig. 1.3. First, we consider our patient's preferences and actions. Second, we consider the healthcare resources available to us. Third, we consider the best available evidence. Fourth, we consider the clinical setting we are practicing (e.g., academic, community, developing nation, war zone), finally we integrate these with our own surgical skills.

In recent years, strides have been made across the globe in the adoption and advancement of EBS. It has become standard for surgical training programs to integrate evidence-based training and research into their core curriculum. As examples in North America, Duke University's surgical departments are associated with the Duke Clinical Research Institute (DCRI). Here, surgical residents can receive training, participate in, and lead clinical research projects [39]. The DCRI strives to create meaningful and realistic clinical trials that can shape patient care, and then integrate the results into practice [39]. Similarly, Stanford University has a Center for Clinical

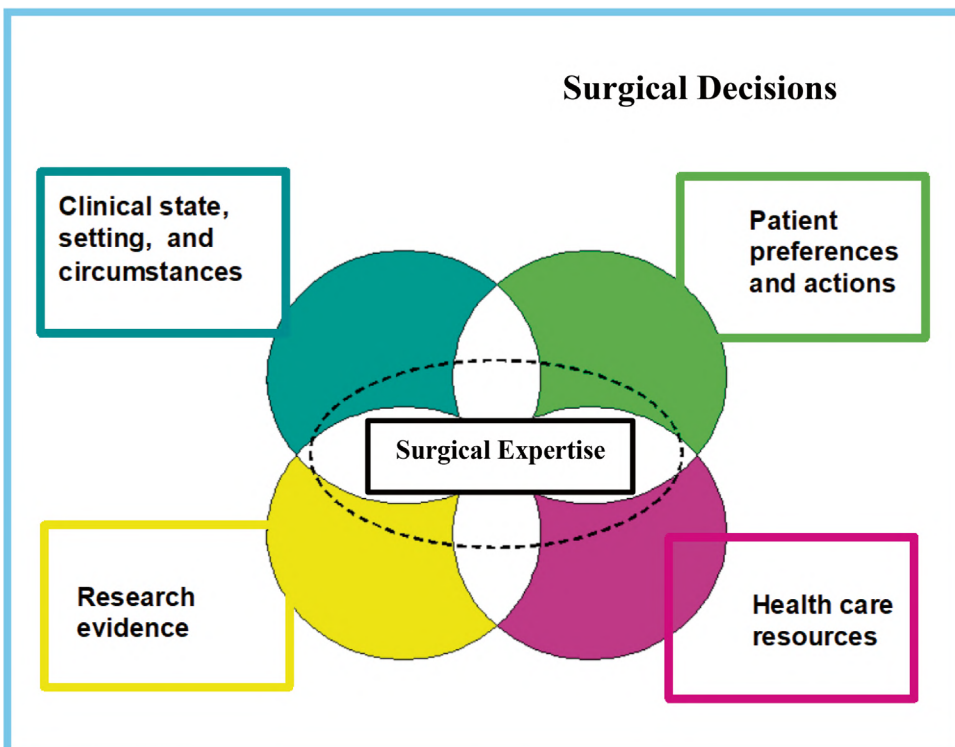


Fig. 1.3 Model for surgical decisions. (Adapted from Haynes et al. [38])

Research (SCCR) that focuses on research collaboration, innovation, and evidence-based operations [40]. The SCCR integrates educational sessions into their program with past sessions on: Clinical Research Education, Good Clinical Practice, and The Informed Consent Process [40]. McMaster University has included very similar programming into their core surgical curriculum by incorporating a surgical research methodology (SRM) course [41]. Through this course, residents learn critical thinking and the appreciation for statistical interpretation [41]. The success of these programs can be shown through the increases in research productivity and performance in surgical residents among all disciplines [40]. It is believed that the implementation of such programs not only benefit the future careers of surgical residents, but also encourage evidence-based patient care [41].

Validated *methods* to evaluate competency in EBM are also being implemented. The FRESNO [42] and Berlin Questionnaire [43] are two such instruments used as metrics in evaluating the success of these programs, allowing for standardized comparison between and across program implementations. Recently, there has been movement toward a nation-based EBS, as opposed to institution-based. In the United Kingdom (UK), the Intercollegiate Surgical Curriculum Programme (ISCP), the approved UK framework for surgical training, now requires that trainees demonstrate use of the best current evidence in decision-making about patients [44].

Since the 1990s, several initiatives have been introduced to improve the methodology and standardize the reporting of various study designs and measurement of outcomes (see Table 1.1). A synopsis of each of these are also discussed below.

One of the earliest initiatives to improve outcome measurement was OMERACT (Outcome Measurements in Rheumatology) [45]. OMERACT is an independent initiative of healthcare professionals that was introduced in the early 1990s [45]. OMERACT continues to develop and validate clinical and radiographic measures in rheumatic diseases [45]. Every two years, OMERACT organizes meetings to develop

outcome measure consensus among panel members within the context of musculoskeletal and autoimmune diseases [45]. Information on the OMERACT group can be found at their website, www.omeract.org.

The introduction of OMERACT was closely followed by GRADE (Grading of Recommendations Assessment, Development and Evaluation). GRADE was developed by an international panel of experts in evidence-based practice [46]. The aim of GRADE was to create a transparent and easy-to-follow guide for rating the quality/certainty of evidence and the strength of recommendations in systematic reviews and guideline development [46]. From this initiative came the development of GRADEpro and its corresponding phone application that can be used as an all-on-one tool for summarizing and presenting information for decision-making in health care [46]. In 2007, the GRADE system was adopted by the World Health Organization (WHO) for the development of public health guidelines [47–49]. It has now become the most widely adopted tool worldwide with over 100 organizations endorsing GRADE for grading the quality of evidence [50]. More information can be found at their website www.grade-workinggroup.org.

Included in those organizations who were early adopters of GRADE, is the Cochrane Collaboration to evaluate quality of evidence from systematic reviews and further, to summarize the findings and present the evidence to decision-makers [51]. The Cochrane Collaboration (named in honor of Archie Cochrane) is an independent network of researchers, professionals, healthcare workers, and patients that was formed in the mid-2000s [52]. Its goal is to improve healthcare decisions and transform the way decisions are made in health care. To accomplish this goal, Cochrane gathers and summarizes the best evidence to help professionals make informed treatment decisions [52]. As of 2024, there are members and supporters from over 190 countries working together to create a source of credible and accessible health information, free of commercial sponsorship or conflict of interest [52]. The contributors include

Table 1.1 Initiatives

Name	Goal	Website
OMERACT (Outcome Measurements in Rheumatology)	Develop and validate clinical and radiographic measures in rheumatic diseases	www.omeract.org
GRADE (Grading of Recommendations Assessment, Development and Evaluation)	Create a transparent and easy-to-follow guide for rating the quality/certainty of evidence and the strength of recommendations in systematic reviews and guideline development	www.gradeworkinggroup.org
Cochrane Collaboration	Improve healthcare decisions and transform the way decisions are made in health care, by gathering and summarizing the best evidence to help professionals make informed treatment decisions	www.cochrane.org
COSMIN (Consensus-based Standards for the selection of health Measurement INstruments)	Regulate and improve the methodological quality of health outcome measurement	www.cosmin.nl
EQUATOR (Enhancing the QUALity and Transparency Of health Research) Network	Improve reliability and value of health research	www.equator-network.org
COMET (Core Outcome Measures in Effectiveness Trials)	Focuses on the development and application of agreed-upon standardized outcome sets or “core outcome sets”	www.comet-initiative.org
clinicaltrials.gov	Monitored database of clinical trials; provides resources on publicly and privately supported clinical trials for patients, family members, and healthcare professionals	www.clinicaltrials.gov/ct2/home
PROSPERO	Provide a comprehensive list of registered systematic reviews to help avoid duplication and reduce the opportunity for reporting bias	www.crd.york.ac.uk/prospero

leaders in the fields of medicine, health policy, research methodology, and consumer advocacy. The work of Cochrane is necessary as access to health evidence, and the subsequent risk of misinterpreting data increases [52]. More information on Cochrane can be found at www.cochrane.org.

To regulate the methodological quality of health measurement, the COSMIN (Consensus-based Standards for the selection of health Measurement INstruments) initiative was developed [53]. In 2006, this group published an appraisal tool for evaluating the quality of studies based on measurement properties of health measurement instruments [53]. The checklist, which was developed in an international Delphi study, focuses on health-related patient-reported outcomes, and is useful in evaluating studies on

performance-based tests or clinical rating scales [53]. More information on the COSMIN initiative, as well as their checklist can be found at www.cosmin.nl.

The EQUATOR (Enhancing the QUALity and Transparency Of health Research) Network is a well-known international initiative that aims to improve reliability and value of health research [53]. To accomplish this goal, the EQUATOR Network has created stringent guidelines to be used for reporting, based on study type [54]. The EQUATOR Network provides information pertaining to all their guidelines on their website www.equator-network.org. As of December 2024, there are 655 reporting guidelines in their database. Table 1.2 lists guidelines for the main study types. Most guidelines have extensions that

can be applied to more specific types of the applicable study designs and even certain clinical areas or section of reports. For example, the extensions for systematic reviews include Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines specific to scoping reviews, diagnostic test accuracy, and abstract reviews [55].

While the abovementioned guidelines from the EQUATOR Network outline proper reporting in the health research literature, the COMET (Core Outcome Measures in Effectiveness Trials) initiative focuses on the development and application of agreed-upon standardized outcome sets or “core outcome sets” [56]. This initiative outlines the minimum that should be both measured and reported within a clinical trial regarding a specific condition [56]. This outlines the expectation that the identified core outcomes will be collected and reported for each respective condition, allowing results from many trials to be compared and combined [56]. The COMET handbook is available at their website: www.comet-initiative.org.

Regulating the quality of both health measurement and health reporting is incredibly important for the standardization of clinical literature. However, maintaining public transparency of methods, anonymized results, and adverse events is potentially just as important for the overall quality of clinical research. In 1997, the Food and Drug Administration (FDA) established ClinicalTrials.gov, an up-to-date and monitored

database of clinical trials [57]. This government-led website, made public in 2000, provides resources on publicly and privately supported clinical trials for patients, family members, and healthcare professionals [57]. Maintained by the National Library of Medicine (NLM) from the National Institutes of Health (NIH), this database is updated and strictly monitored to ensure up-to-date information [57]. This registry is a government website and therefore does not host, receive funding, or advertise any commercial entities or content [57]. As of December 2024, there are over 519,000 study records on the ClinicalTrials.gov database that are being conducted in all 50 States and 223 countries [57]. The registry is a great way to stay current on developments in the various fields of medicine, find collaborating partners and identify gaps within specific areas of interest. Anyone can search this database, without registering, by visiting www.clinicaltrials.gov/ct2/home.

Researchers can also register systematic reviews with PROSPERO, the International prospective registrar of systematic reviews [58]. PROSPERO was produced by the Centre for Reviews and Dissemination at York University and is funded by the NIH [58]. This international database contains detailed information about systematic review protocols in numerous areas including health and social care, welfare, public health, and education [58]. Key features from each review protocol are recorded and maintained to provide a comprehensive list of regis-

Table 1.2 Reporting guideline by study type

Study type	Suggested guideline(s)
Randomized trials	CONSORT (+ Extensions)
Observational studies	STROBE (+ Extensions)
Systematic reviews	PRISMA (+ Extensions)
Case reports	CARE (+ Extensions)
Qualitative research	SRQR/COREQ
Diagnostic/prognostic studies	STARD/TRIPOD
Quality improvement studies	SQUIRE (+ Extensions)
Economic evaluations	CHEERS (+ Extensions)
Animal preclinical studies	ARRIVE
Study protocols	SPIRIT/PRISMA-P
Clinical practice guidelines	AGREE/RIGHT

Source: The EQUATOR Network [54]

tered systematic reviews to help avoid duplication and reduce the opportunity for reporting bias [58]. Cochrane protocols are also included in PROSPERO with links to the full protocol found on the Cochrane Library. Readers can learn more about PROSPERO by visiting: www.crd.york.ac.uk/prospero.

The application of the abovementioned guidelines can help to address the many areas of failure that can occur during the research process [59]. The EQUATOR Network guidelines, the COMET handbook, and the COSMIN tool are critical in the design, implementation, and reporting of clinical and observational trials; GRADE can be applied in systematic reviews, meta-analyses, and guideline development [59]. Through the integration of the recommendations of the EQUATOR guidelines, surgeons will advance reporting and transparency of surgical research; similarly, by applying COSMIN, surgical research will be more efficient. These initiatives will thus advance the overall philosophy of EBS. A future challenge to EBS that needs to be addressed is Knowledge Translation, moving novel research findings into surgical practice. With the plentiful publication of new information, the “true signal” of surgical advances may be lost in the “noise.” The new generation of surgeons should be trained in appraisal skills and informed of important pre-appraised information such as systematic reviews or clinical practice guidelines. Utilizing the EQUATOR Network, COMET, COSMIN, and GRADE guidelines can help surgeons appraise and be aware of clinically important literature within their subspecialties.

References

1. Guyatt GH. Evidence-based medicine. *ACP J Club*. 1991;A-16:114.
2. Smith R, Rennie D. Evidence based medicine—an oral history. *BMJ*. 2014;348:371–3.
3. Bhatt A. Evolution of clinical research: a history before and beyond James Lind. *Perspect Clin Res*. 2010;1(10):6–10.
4. Imber G. *Genius on the edge: the bizarre double life of Dr. William Steeward Halsted*. New York: Kaplan Publishing; 2011.
5. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. *Lancet*. 2017;290(10092):415–23.
6. Meakins JL. Innovation in surgery: the rules of evidence. *Am J Surg*. 2002;183(4):399–405.
7. Canadian Task Force on the Periodic Health Examination (CTFPHE). The periodic health examination. *Can Med Assoc J*. 1979;121(9):1193–254.
8. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest*. 1989;95(Suppl 2):2S–4S.
9. Eddy DM, Billings J. The quality of medical evidence: implications for quality of care. *Health Aff (Millwood)*. 1988;7:19–32.
10. Cochrane AL. *Effectiveness and efficiency: random reflections on health services*. London: Nuffield Provincial Hospitals Trust; 1972.
11. Eddy DM. Guidelines for the cancer related checkup: Recommendations and rationale. *CA Cancer J Clin*. 1980;30:4–50.
12. Sackett DL. How to read clinical journals: I. Why to read them and how to start reading them critically. *Can Med Assoc J*. 1981;124(5):555–8.
13. Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. The evidence-based medicine working group. *JAMA*. 1993;270(17):2093–5.
14. Oxford Centre for Evidence-based Medicine: Levels of Evidence (March 2009) [Internet]. Centre for Evidence-Based Medicine [cited 2024 Apr 01]. Available from <https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009>
15. Archibald S, Bhandari M, Thoma A. Evidence based surgery series: users' guide to the surgical literature: How to use an article about a diagnostic test. *Can J Surg*. 2001;44(1):17–23.
16. Urschel J, Tandan V, Miller J, Goldsmith CH. Users' guide to evidence-based surgery: how to use an article evaluating surgical interventions. *Can J Surg*. 2001;44(2):95–100.
17. Thoma A, Sprague S, Tandan V. Users' guide to the surgical literature: how to use an article on economic analysis. *Can J Surg*. 2001;44(5):347–55.
18. Hong D, Tandan V, Goldsmith CH, Simunovic M. Users' guide to the surgical literature: how to use an article reporting population-based volume-outcome relationships in surgery. *Can J Surg*. 2002;45(2):109–15.
19. Birch D, Eady A, Robertson D, De Pauw S, Tandan V. Users' guide to the surgical literature: How to perform a literature search. *Can J Surg*. 2003;46(2):136–41.
20. Bhandari M, Devereaux PJ, Montori V, Cinà C, Tandan V, Guyatt GH. Users' guide to the surgical literature: How to use a systematic literature review and meta-analysis. *Can J Surg*. 2004;47(1):60–7.
21. Thoma A, Farrokhyar F, Bhandari M, Tandan V, The Evidence-Based Surgery Working Group. Users' guide to the surgical literature: How to assess a

- randomized controlled trial in surgery. *Can J Surg.* 2004;47(3):200–8.
22. Birch D, Goldsmith CH, Tandan V. Users' guide to the surgical literature: self audit and practice appraisal for surgeons. *Can J Surg.* 2005;48(1):57–62.
23. Mastracci TM, Thoma A, Farrokhyar F, Tandan VR, Cina CS. Users' guide to the surgical literature: how to use a decision analysis. *Can J Surg.* 2007;50(5):403–9.
24. Thoma A, Cornacchi S, Lovrics P, Goldsmith CH. Users' guide to the surgical literature: how to assess an article on health-related quality of life. *Can J Surg.* 2008;51(3):215–24.
25. Cadeddu M, Farrokhyar F, Thoma A, Haines T, Garnett A, Goldsmith CH. Users' guide to the surgical literature: how to assess power and sample size. *Can J Surg.* 2008;51(6):476–82.
26. Hansebout R, Cornacchi SD, Haines T, Goldsmith CH. Users' guide to the surgical literature: how to use an article on prognosis. *Can J Surg.* 2009;52(4):328–36.
27. Dijkman B, Kooistra B, Bhandari M. Users' guide to the surgical literature: how to work with a subgroup analysis. *Can J Surg.* 2009;52(6):515–22.
28. Thoma A, Cornacchi SD, Farrokhyar F, Bhandari M, Goldsmith CH. Users' guides to the surgical literature: how to assess a survey in surgery. *Can J Surg.* 2011;54(6):394–402.
29. Cadeddu M, Farrokhyar F, Levis C, Cornacchi HT, Thoma A, Evidence-Based Surgery Working Group. Users' guide to the surgical literature: Understanding confidence intervals. *Can J Surg.* 2012;55(3):207–11.
30. Coroneos CJ, Voineskos SH, Cornacchi SD, Goldsmith CH, Ignacy TA, Thoma A. Users' guide to the surgical literature: how to evaluate clinical practice guidelines. *Can J Surg.* 2014;57(4):280–6.
31. Waltho D, Kaur MN, Haynes RB, Farrokhyar F, Thoma A. Users' guide to the surgical literature: how to perform a high-quality literature search. *Can J Surg.* 2015;58(5):349–58.
32. Thoma A, Kaur MN, Farrokhyar F, Waltho D, Levis C, Lovrics P, Goldsmith CH. Users' guide to the surgical literature: how to assess an article harm in surgery. *Can J Surg.* 2016;59(5):351–7.
33. Gallo L, Eskicioglu C, Braga L, Farrokhyar F, Thoma A. Users' guide to the surgical literature: how to assess an article using surrogate outcomes. *Can J Surg.* 2017;60(4):280–7.
34. Thoma A, Farrokhyar F, Waltho D, Braga LH, Sprague S, Goldsmith CH. Users' guide to the surgical literature: how to assess a noninferiority trial. *Can J Surg.* 2017;60(6):426–32.
35. Gallo L, Murphy J, Braga L, Farrokhyar F, Thoma A. Users' guide to the surgical literature: how to assess a qualitative study. *Can J Surg.* 2018;61(3):208–14.
36. Thoma A. Evidence-based plastic surgery: design, measurement and evaluation. *Clin Plastic Surg.* 2008;35(2):189–316.
37. Bhandari M, editor. Evidence-based orthopedics. Oxford: Wiley-Blackwell; 2012.
38. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. *BMJ.* 2002;7:36–8.
39. Duke Clinical Research Institute: Designing Studies That Unveil Truths [Internet]. Duke University [cited 2024 Apr 01]. Available from <https://dcric.org/about>
40. Stanford University. Stanford Center for Clinical Research (SCCR) [Internet]. Stanford University; c2018 [cited 2024 Apr 01]. Available from <http://med.stanford.edu/sccr.html>
41. Farrokhyar F, Amin N, Dath D, Bhandari M, Kelly S, Kolkin AM, Gill-Pottruff C, Skot M, Reid S. Impact of the surgical research methodology program on surgical residents' research profiles. *J Surg Educ.* 2014;71(4):513–20.
42. Ramos KD, Schafer S, Tracz SM. Validation of the Fresno test of competence in evidence based medicine. *BMJ.* 2003;326(7384):319–21.
43. Fritsche L, Greenhalgh T, Falck-Ytter Y, Neumayer HH, Kunz R. Do short courses in evidence based medicine improve knowledge and skills? Validation of Berlin questionnaire and before and after study of courses in evidence based medicine. *BMJ.* 2002;325(7376):1338–41.
44. Intercollegiate Surgical Curriculum Programme. The syllabus [Internet]. United Kingdom: JCST [cited 2024 Apr 01]. Available from https://www.iscp.ac.uk/curriculum/surgical/surgical_syllabus_list.aspx
45. OMERACT Outcome Measures in Rheumatology. Welcome to OMERACT [Internet]. OMERACT [cited 2024 Apr 01]. Available from <http://www.omeract.org>
46. GRADE Working Group. GRADE: Welcome to the GRADE Working Group – From Evidence to recommendations – transparent and sensible [Internet]. GRADE Working Group [cited 2024 Apr 01]. Available from <http://www.gradeworkinggroup.org/>
47. Alexander PE, Brito JP, Neumann I, Gionfriddo MR, Bero L, Djulbegovic B, Stoltzfus R, Montori VM, Norris SL, Schünemann HJ, Guyatt GH. World Health Organization strong recommendations based on low-quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. *J Clin Epidemiol.* 2016;72(Complete):98–106.
48. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ, GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ.* 2008;336(7650):924–6.
49. WHO Handbook for Guideline Development [Internet]. World Health Organization [cited 2024 Jun 06]. Available from http://apps.who.int/iris/bitstream/10665/75146/1/9789241548441_eng.pdf
50. BMJ Best Practice. What is GRADE? [Internet]. BMJ [cited 2024 Jun 06]. Available from <https://bestpractice.bmj.com/info/us/toolkit/learn-ebm/what-is-grade/>

51. McMaster University Mac GRADE Centre. About GRADE [Internet]. McMaster University [cited 2024 Apr 01]. Available from <https://cebgrade.mcmaster.ca/aboutgrade.html>
52. Cochrane. About Us [Internet]. Cochrane Group [cited 2024 Apr 01]. Available from <http://www.cochrane.org/about-us>
53. VU University Medical Center. COSMIN [Internet]. VU University [cited 2024 Apr 01]. Available from <http://www.cosmin.nl/>
54. The EQUATOR Network. equator network—About Us [Internet]. The EQUATOR Network [cited 2024 Dec 12]. Available from <https://www.equator-network.org/about-us/>
55. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffman TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hrobjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: An updated guidelines for reporting systematic review. *BMJ*. 2021;372:n71.
56. COMET Initiative. The COMET Initiative [Internet]. COMET [cited 2024 Apr 01]. Available from <http://www.comet-initiative.org/>
57. National Institute of Health (NIH). [Clinicaltrials.gov](https://www.clinicaltrials.gov) [Internet]. NIH [cited 2024 Dec 12]. Available from <https://www.clinicaltrials.gov/>
58. University of York Centre for Reviews and Dissemination—PROSPERO. About PROSPERO [Internet]. York [cited 2024 Apr 01]. Available from <https://www.crd.york.ac.uk/prospERO/#aboutpage>
59. Innes N, Schwendicke F, Lamont T. How do we create, and improve the evidence? *Br Dent J*. 2016;220(12):651–5.

The Steps of Practicing Evidence-Based Surgery (EBS)

2

Achilles Thoma, Emily Dunn, Sheila Sprague,
Luis H. Braga, and Sophocles H. Voineskos

Abbreviations

EBM	Evidence-Based Medicine
EBS	Evidence-Based Surgery
PICOT	Population Intervention Comparison Outcome Time Horizon
OR	Odds Ratio
CI	Confidence Interval

In Chap. 1: *The History of Evidence-Based Surgery*, we explained the historical development of the evidence-based medicine (EBM) and evidence-based surgery (EBS) and paradigm shifts of practicing medicine and surgery to the present era. EBS is defined as an approach of

practicing surgery in which the surgeon is aware of the best evidence in support of practice and the strength of that evidence.

In 2008, the *British Medical Journal (BMJ)* hailed this paradigm shift as one of the 15 most important milestones in medicine in the last 150 years [1]. The experiential and authoritarian approach to learning and practicing medicine and surgery in the past has been supplanted by the evidence-based approach [2]. Many surgeons, however, are not familiar with the concepts of EBS and how they can be applied. We hope that this book will help close that gap.

This particular chapter will show you step by step how to adopt an evidence-based approach to

A. Thoma (✉)
Division of Plastic Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: athoma@mcmaster.ca

E. Dunn
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: dunne4@mcmaster.ca

S. Sprague
Division of Orthopaedic Surgery, Department of
Surgery & Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: sprags@mcmaster.ca

L. H. Braga
Department of Health Research Methods, Evidence,
and Impact, McMaster University,
Hamilton, ON, Canada

Division of Urology, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: braga@mcmaster.ca

S. H. Voineskos
Division of Plastic, Reconstructive, and Aesthetic
Surgery, Department of Surgery, University of
Toronto, & Li Ka Shing Knowledge Institute,
Unity Health, Toronto, ON, Canada
e-mail: sophocles.voineskos@medportal.ca

Table 2.1 Steps to evidence-based surgery

1	Construct a relevant, answerable <i>question</i> from a clinical scenario (see <i>Chap. 3: Developing a Surgical Research Question</i>)
2	Plan and carry out a literature <i>search</i> for best external evidence (see <i>Chap. 4: Finding the Evidence through Searching the Literature</i>)
3	Critically <i>appraise</i> the literature for validity and applicability (see <i>Chap. 6: Evaluating Surgical Interventions, Chap. 10: How to Assess an Article that Deals with Health-Related Quality of Life, and Chap. 11: Randomized Controlled Trial Comparing Surgical Interventions</i>)
4	Apply the evidence to your clinical practice (see <i>resolution of scenarios in Chaps. 10, 11, 24</i>)
5	Evaluate your performance

your surgical practice. There are five distinct steps in the process of systematically finding, appraising, and using the contemporaneous research findings as the basis for clinical decisions [3].

These steps need to be followed precisely in their correct order (see Table 2.1). They will become clearer as you read the following chapters of the book, which go into more detail and also explain the methodological issues of the various study designs.

Step 1 Construct a relevant, answerable *question* from a clinical scenario.

The construction of a clinically relevant, answerable question from a surgical case that you encounter in your clinical practice, is the first step. This step is important for both the “Users” of clinical research, that is, those who want to know, like most of us surgeons but also other stakeholders such as other clinicians, patients, patients’ families, policy decision-makers, administrators, and general public. It is also necessary for the “Doers,” such as investigator surgeons, who intend to find the answer to a surgical question through clinical research.

An example of a relevant clinical question is described below.

Scenario

At the weekly general surgery rounds, a senior surgeon challenged a newly appointed colorectal surgeon when they claimed that the incidence of ureteral injury is lower with a laparoscopic

approach rather than with open surgery because of better visibility.

This is a plausible scenario that can generate enough controversy to create a research question. The answer to this question may not be readily available. To find it, we proceed as suggested in Step 1, developing a clinically relevant and answerable question from this scenario. The question is structured in such a way as to identify the *Population or Patients, Intervention, Comparative Intervention, Outcome, and Time Horizon*. These terms can be remembered by the acronym *PICOT* [4, 5].

For the abovementioned scenario, the proper research question using the structure of PICOT would be, “In colorectal cancer patients, does laparoscopic, as compared to open surgery, result in less ureteral injury 30 days following surgery?”

As a rule of thumb, we ask such questions in one breath. If you cannot ask such PICOT formatted question in one breath, it means that your question is convoluted. Try practicing “one breath” PICOT questions until you become an expert.

The structure and formulation of the research question is discussed in more detail in *Chap. 3: Developing a Surgical Research Question*.

Step 2 Plan and carry out a *search* of the literature for best external evidence.

For the literature search conducted in this step, we select keywords using our PICOT formulation, which, as you remember is generated from our clinical scenario/problem. Such keywords will include “colon cancer”, “laparoscopic surgery,” “open surgery,” and “ureter injury.”

These keywords are then first entered into a filtered electronic database to see if any articles address the research question. Filtered databases, such as the Cochrane Database, include articles pre-appraised by methodological experts. If there are no hits obtained on this database, an unfiltered database could be used. Unfiltered databases are ones in which articles go through the usual review process; examples of unfiltered database are Medline or Embase.

The literature search is an important step, and all surgeons should be familiar with it; for more detailed information on how to perform and evaluate a literature search, please see *Chap. 4: Finding the Evidence Through Searching the Literature*.

Step 3 Critically *appraise* the literature for validity and applicability.

The appraisal of a surgical article involves answering three important questions:

Question #1: Is the study presented in the article valid?

Question #2: If valid, what were the results?

Question # 3: Can the results be applied to my patients or practice?

Depending on the study design we are examining, each of these three main questions may have additional subsidiary questions. These questions will be a recurrent theme in the chapters to follow. If on reading the article, particularly the methods section, you do not believe the study is valid, do not waste your time reading the article in its entirety. If on the other hand, you believe it is valid then you should proceed and examine the results and what they may mean for the study group in the article and potentially, your own patients. *Chapter 6: Evaluating Surgical Interventions* explains how to interpret the results of surgical interventions. This book contains chapters dedicated to explaining how to appraise properly, the evidence, based on the specific content or methodology used in a study.

Step 4 *Apply* the evidence to your clinical practice.

If you believe that the investigators have used correct methods in their study and on assessing their results, you find that the novel intervention shows better results than the standard approach you are using, then you need to decide if you can apply them to your practice.

This decision will be made if you believe that the patients examined in the article you read are similar to your patients. If for example, the study was done in pediatric patients, but your patients are adults, the study's results may not be applicable to your practice. You may also consider the clinical resources available to you as you may be working in a rural hospital, where you may not have access to the latest technological advances (e.g., robotic surgery, single-port laparoscopic instrumentation).

Step 5 *Evaluate* your effectiveness and efficiency in carrying out Steps 1–4.

- A. Evaluate your performance in asking answerable questions
- B. Evaluate your performance in searching and finding external evidence through electronic database searching
- C. Evaluate your performance in critical appraisal
- D. Evaluate your performance in integrating evidence and patients' values into your practice

Make sure to conduct an ongoing evaluation, ideally with the help of a mentor, of the effects the implementation of this new evidence has made in your practice.

How to Apply Steps 1–5:

An example of how to apply the abovementioned steps into your practice is shown below. For this example, the following scenario will be used:

At the weekly general surgery rounds, a senior surgeon challenged a newly appointed colorectal surgeon (who just completed their fellowship) when the young surgeon claimed that the incidence of ureteral injury is lower

with the laparoscopic technique than the open technique during colorectal surgery. They claimed that the laparoscopic method provides better visualization and thus is less likely to lead to damage of the ureter. The head of the Division asks the Minimal Access Surgery Fellow in the Service to review the surgical literature and report whether this is true or not at next week's rounds.

Step 1 Construct a relevant, answerable question from a clinical scenario.

Using the PICOT as a guide, your clinical question would include the following:

Population: Patients undergoing colorectal surgery

Intervention: Laparoscopic surgery

Comparison Intervention: Open-technique surgery

Outcome: Ureteral injury

Time Horizon: Up to 30 days following surgery

Therefore, the finalized research question would be: *In patients undergoing colorectal surgery, is laparoscopic surgery less likely to result in damage to the ureter than open-technique surgery, 30 days postsurgery?*

Step 2 Plan and carry out a search of the literature for best external evidence.

Using your search tool of choice, whether filtered, or unfiltered, the PICOT elements would be entered in order to find the best evidence for your question. For example, to find articles related to the above-stated research question, we would want to use the following search terms: *Colorectal Cancer AND Ureteral Injury AND Laparoscopic Surgery AND Open surgery.*

Below is an example of how you would perform a literature search using The Cochrane Library (Fig. 2.1a, b).

As an appropriate article was not found using the Cochrane database, a search in PubMed can be used (Fig. 2.2a, b). We limited the search to within the last 10 years: 2013–April 2024.

Using PubMed, using the same search terms, 13 articles, which were better suited to answer the research question, were found.

In the first edition of this textbook, published in 2019, we applied the same search strategy to answer the research question. At that time, the search identified five articles, and the article “Iatrogenic ureteral injury in colorectal cancer surgery: A nationwide study comparing laparoscopic and open approaches” by Andersen et al. in 2015 [6], was chosen as the most applicable. In this updated search, Andersen et al. [6], is again identified. However, newer articles also appear, including meta-analyses. You select one titled: “Iatrogenic ureteric injury during abdominal or pelvic surgery: A meta-analysis” by Yanagisawa et al. [7] published in 2023. This article appears to be a possibly relevant article to give answers to your research question. You then proceed to read the abstract shown in Fig. 2.3, which continues to look promising, so we then proceed to read the whole article. The article presents a comprehensive systematic review and several meta-analyses, in which the Andersen et al. [6] article was included. We will focus on the portion of the article regarding ureteric injury incidence in laparoscopic vs open surgery.

Step 3 Critically appraise the literature for validity and applicability.

Question 1: Is the study valid?

The validity of the chosen study will be judged in different ways based on the methodology and focus of the article. For example, with this article, we may look at search strategy. Yanagisawa et al. [7] state that they followed the PRISMA guidelines and include the PRISMA checklist in supplementary files. Also included in the supplementary files is the full search strategy, allowing for reproducibility.

Secondly, to see if they addressed the quality of the included articles we can check if the authors performed a risk of bias assessment. In this case, the authors applied the risk of bias in non-randomized studies of interventions (ROBINS-I). As such, the heterogeneity of the

Fig. 2.1 (a) Search terms from your research questions are entered. (b) A list of appropriate articles

For the results of the Yanagisawa et al. [7] article, one could look at the association between

the exposed and the outcome. When pooling the data from the 16 included studies, there were 243 (0.43%) ureteric injuries reported in the patients who underwent laparoscopic colectomy and 665 (0.33%) in the patients who underwent open colectomy. The analysis revealed no significant difference in the incidence of ureteric injuries between laparoscopic and open colectomy

a



National Library of Medicine
National Center for Biotechnology Information

Log in

PubMed Advanced Search Builder


User Guide

Add terms to the query box

All Fields

Enter a search term

AND

Show Index

Query box

(((Colorectal Cancer) AND (Ureteral Injury)) AND (Laparoscopic surgery)) AND (Open Surgery)

Search

Fig. 2.2 (a) Search terms from your research question are entered. (b) A list of appropriate articles for your research question is listed

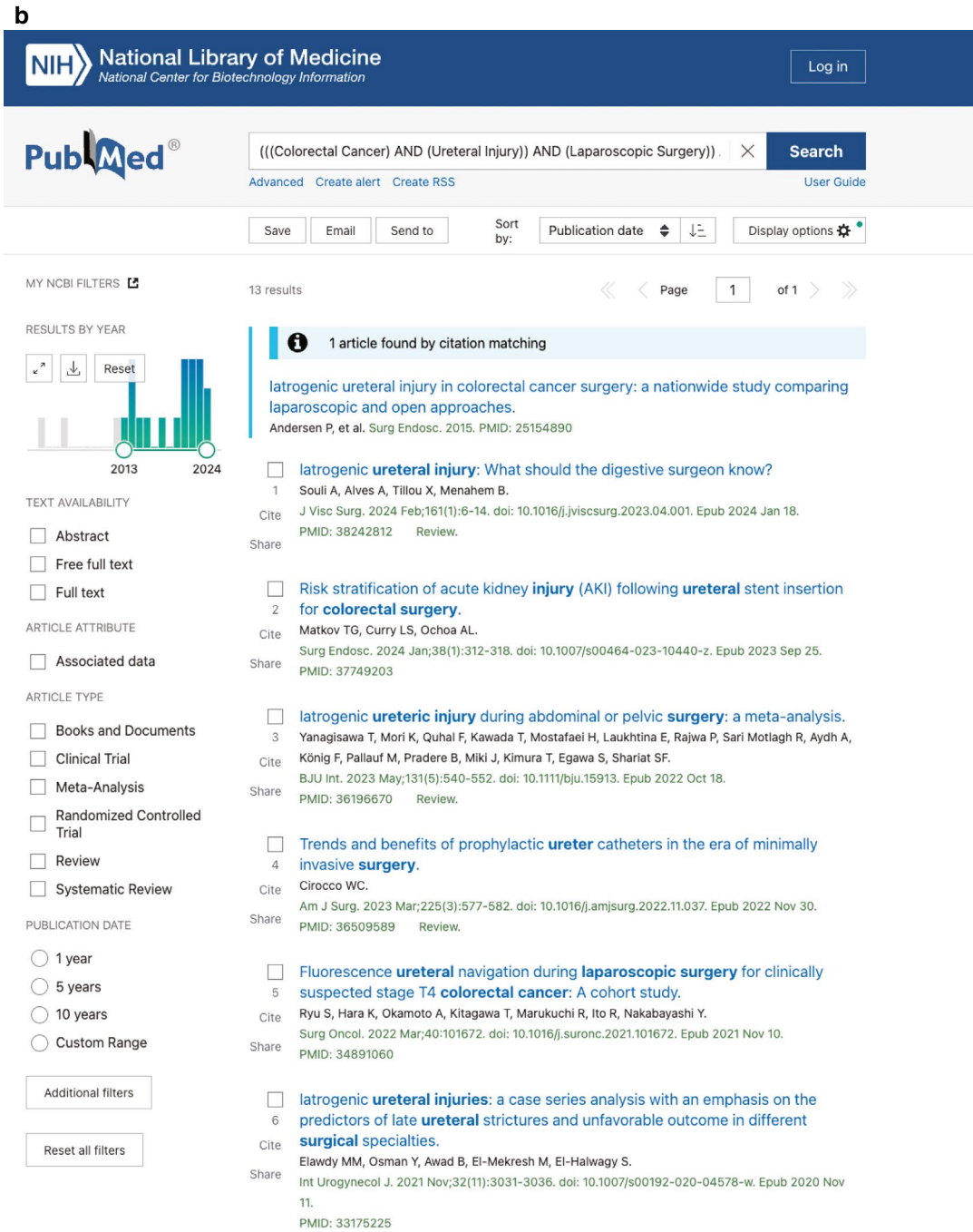


Fig. 2.2 (continued)

Review > BJU Int. 2023 May;131(5):540-552. doi: 10.1111/bju.15913. Epub 2022 Oct 18.

Iatrogenic ureteric injury during abdominal or pelvic surgery: a meta-analysis

Takafumi Yanagisawa^{1,2}, Keiichiro Mori², Fahad Qahal^{1,3}, Tatsushi Kawada^{1,4}, Hadi Mostafaei^{1,5}, Ekaterina Laukhtina¹, Pawel Rajwa^{1,6}, Reza Sari Motlagh^{1,7}, Abdulmajeed Aydh^{1,8}, Frederik König^{1,9}, Maximilian Pallauf^{1,10}, Benjamin Pradere¹, Jun Miki², Takahiro Kimura², Shin Egawa², Shahrokh F Shariat^{1,11,12,13,14,15}

Affiliations + expand

PMID: 36196670 DOI: 10.1111/bju.15913

Abstract

Objective: To assess the incidence of ureteric injuries, clinical value of prophylactic ureteric stenting and impact of intra- or postoperative detection of ureteric injuries in patients treated with gynaecological or colorectal surgery.

Methods: Multiple databases were searched for articles published before September 2021 according to the Preferred Reporting Items for Systematic Review and Meta-Analyses statement. Studies were deemed eligible if they evaluated the differences in the rate of ureteric injuries between laparoscopic and open surgery, prophylactic ureteric stenting or not, and those of final treatment success between intra- and postoperative detection in patients who underwent gynaecological or colorectal surgery.

Results: Overall, 46 studies were eligible for this meta-analysis. Compared to open surgery, laparoscopic hysterectomy was associated with a higher incidence of ureteric injuries (pooled odds ratio [OR] 2.12, 95% confidence interval [CI] 1.71–2.62), but there was no statistically significant difference in colectomy (pooled OR 0.89, 95% CI 0.77–1.03). Prophylactic ureteric stenting was associated with a lower incidence of ureteric injuries during gynaecological surgery (pooled OR 0.61, 95% CI 0.39–0.96). The number needed to perform ureteric stenting to prevent one ureteric injury was 224 in gynaecological surgery. On the other hand, prophylactic ureteric stenting did not reduce the risk of ureteric injuries during colorectal surgery. Intraoperative detection of a ureteric injury was associated with a lower rate of complication management failure compared to postoperative detection (pooled OR 0.22, 95% CI 0.12–0.41).

Conclusions: Laparoscopic hysterectomy seems to be associated with a higher rate of ureteric injuries compared to an open approach. Prophylactic ureteric stenting seems to reduce this risk during gynaecological surgery. Intraoperative detection of a ureteric injury during abdominal/pelvic surgery improves outcomes, suggesting the need for awareness and proactive problem identification. Further well-designed studies assessing the candidates who are more likely to benefit from prophylactic ureteric stenting including cost analysis are needed.

Keywords: #UroTrauma; #Urology; colectomy; hysterectomy; iatrogenic; laparoscopic; ureteric injury; ureteric stent.

© 2022 BJU International.

FULL TEXT LINKS

WILEY Full Text Article

ACTIONS

Cite

Collections

SHARE



PAGE NAVIGATION

< Title & authors

Abstract

Similar articles

Cited by

References

Publication types

MeSH terms

Related information

LinkOut - more resources

Fig. 2.3 Abstract review of chosen article

(pooled odds ratio (OR) 0.89, 95% confidence interval (CI) 0.77–1.03, $P = 0.11$). However, there was significant heterogeneity, mostly caused by the Andersen et al. [6] study. Whereas the Andersen et al. [6] study concluded that there was higher risk of iatrogenic injury to the ureter with the laparoscopic approach with an OR of 1.64 (95% CI 1.02–2.63, $P = 0.04$) the Yangisawa et al. [7] meta-analysis showed no difference. Yangisawa et al. [7] claimed that laparoscopic surgery was not significantly associated with an

increased risk of iatrogenic ureteral injury (an odds ratio greater than 1 would indicate a higher risk of harm in the exposed group).

Question 3: Can the results be applied to my patients or practice?

When answering question three, one needs to consider if the patients in the study were similar to patients in your practice, if the exposure is similar to what might occur in your patients, the magni-

tude of the risk and any possible benefits associated with the exposure. In this example article, the patients in the study were from a variety of countries, therefore, this may be something to consider when applying results to your patients. Furthermore, this study included more than just colorectal surgeries for cancer, it also included diverticular disease or inflammatory bowel disease. The Andersen et al. [6] study was the only included study that looked at colorectal cancer alone. However, they mention in their discussion section that the learning curve phenomenon could be a factor in the results, as their study was completed during an implementation period for laparoscopic colorectal cancer across Denmark.

Step 4 *Apply the evidence to your clinical practice.*

If you believe that the patients and the technique and magnitude of risk and benefits described in the article are all similar to yours, you should consider applying the evidence to your practice.

Step 5 *Evaluate your performance.*

The final step in EBS practice is the continuous assessment of your performance in implementing all the steps mentioned above. You may want to approach a mentor and ask him/her to provide you with feedback on whether you mastered the skills of (a) constructing answerable

questions on the PICOT format, (b) performing successful literature searches to find the best evidence, (c) appraising what you read and are capable of separating “the chaff from the wheat,” and (d) implementing what you found in your practice.

References

1. Kamerow D. Milestones, tombstones, and sex education. *BMJ*. 2007;334.
2. Meakins JL. Innovation in surgery: the rules of evidence. *Am J Surg*. 2002;183(4):399–405.
3. Sackett DL, Strauss SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. 2nd ed. Edinburgh, Scotland: Churchill Livingstone; 2000.
4. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. *ACP J Club*. 1995;123(3):A12–3.
5. Nollan R, Finout-Overhold E, Stephenson P. Asking compelling clinical questions. In: Melnyk BM, Fineout-Overhold E, editors. Evidence-based practice in nursing and healthcare: a guide to best practice. 2nd ed. Philadelphia: Lippincott Williams & Wilkins; 2005.
6. Andersen P, Andersen LM, Iversen LH. Iatrogenic ureteral injury in colorectal cancer surgery: a nationwide study comparing laparoscopic and open approaches. *Surg Endosc*. 2015;29(6):1406–12.
7. Yanagisawa T, Mori K, Quhal F, Kawada T, Mostafaei H, Laukhtina E, Rajwa P, Sari Motlagh R, Aydh A, König F, Pallauf M, Pradere B, Miki J, Kimura T, Egawa S, Shariat SF. Iatrogenic ureteric injury during abdominal or pelvic surgery: a meta-analysis. *BJU Int*. 2023;131(5):540–52.

Developing a Surgical Clinical Research Question: *To Find the Answer in Literature Search or in Pursuing Clinical Research*

Achilles Thoma, Emily Dunn, Sheila Sprague,
Sophocles H. Voineskos, and Jessica Murphy

Abbreviations

FINER Criteria	F: Feasible; I: Interesting; N: Novel; E: Ethical; R: Relevant
PICOT Format	P: Population; I: Intervention; C: Comparison Intervention; O: Outcome; T: Time Horizon
QoL	Quality of Life
DIEP	Deep Inferior Epigastric Perforator
NIH	National Institute of Health
TAP Blocks	Transverse Abdominal Plane Blocks
ECTR	Endoscopic Carpal Tunnel Release
OCTR	Open Carpal Tunnel Release

In clinical practice, we are faced with clinical questions daily by our medical students, surgical residents, other learners, colleagues, or patients. While we frequently know, and can readily provide an answer, sometimes this is not the case, and we must proceed to find it through a literature search. In such cases, we suggest utilizing the approach to literature searches described by Banfield et al., in Chap. 4. If we do not know the answer and we cannot find it in the surgical literature, we may decide to find it by performing a clinical research project.

Whether it is to find the evidence through a literature search or via a clinical experimentation, the structure of the research question follows the same format. This chapter will explain how to

A. Thoma (✉)
Division of Plastic Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: athoma@mcmaster.ca

E. Dunn
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: dunne4@mcmaster.ca

S. Sprague
Division of Orthopaedic Surgery, Department of
Surgery & Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: sprags@mcmaster.ca

S. H. Voineskos
Division of Plastic, Reconstructive, and Aesthetic
Surgery, Department of Surgery, University of
Toronto, & Li Ka Shing Knowledge Institute,
Unity Health, Toronto, ON, Canada
e-mail: sophocles.voineskos@medportal.ca

J. Murphy
Division of Physical Medicine & Rehabilitation,
Department of Medicine, McMaster University,
Hamilton, ON, Canada
e-mail: murphj11@mcmaster.ca

properly form a research question, as it is the foundation of clinical evidence and proper clinical research. To understand and interpret findings of a surgical article, or to design one's own clinical research project, a surgeon must master the art of formulating "the clinical research question". The posing of the research question should be automatic in a surgeon's daily practice just like driving. The same applies for the surgeon-investigator; however, in this case, prior knowledge on the subject must be identified through a literature search. Being familiar with the current knowledge in an area is also known as understanding the "boundary of knowledge": this must be done prior to forming the research question [1].

There are two clinical questions frequently asked in surgery: (1) background and (2) foreground questions [2, 3]. The *background questions* are usually asked by medical students, surgical interns, and other learners early in their training. These types of questions ask for general knowledge about a surgical problem and have two essential components: (1) a question root (what, who, when, where, how, why) with a verb and (2) a surgical problem. Examples of *background questions* are: (1) What causes ileus after laparotomy?, (2) Why do some mastectomy skin flaps undergo necrosis?, and (3) When should a temporary ileostomy be reversed?

The foreground questions, however, should be asked by senior surgical residents, fellows, and surgeons. These questions ask about the management of patients with surgical problems.

It is important for a research question to be persuasive and clear. A compelling and concise research question helps guide and create a suc-

cessful study and is appealing to both grant agencies and institutions.

This chapter will focus on two tools that can be used to help form a proper research question: (1) the FINER (Feasible, Interesting Novel, Ethical, Relevant) Criteria [4–6] and (2) the PICOT (Population, Intervention, Comparative intervention, Outcomes, Time Horizon) formulation [7, 8].

The FINER Criteria

The FINER criteria (Table 3.1) have been proposed as important preconditions to develop a good research question [4–6]. These criteria, outlined further below, help to highlight the aspects that should be addressed to increase the chances of success in a research project.

Feasible

This aspect focuses on whether there are (1) an adequate number of patients (participants), (2) technical expertise, (3) resources (time, support staff, and money), and (4) a proper study design to make the project manageable [4–6]. There are numerous barriers that may arise with each research project.

For example, a study with the following research question: "*Does the supramicrosurgical periumbilical abdominal flap provide better quality of life (QoL) as compared to the deep inferior epigastric perforator (DIEP) flap in breast reconstruction after mastectomy?*" may experience the following technical barriers: (1)

Table 3.1 The FINER criteria

Criteria	Definition
<i>Feasible</i>	Is the research question manageable? (i.e., enough participants, expertise, time, and money)
<i>Interesting</i>	Does the investigator and the target audience find the research question to be clinically important?
<i>Novel</i>	Does the research question contribute new information to the literature?
<i>Ethical</i>	Is the question ethical? (a clinical research question should be investigated when it is ethical to do so; it requires approval by a research ethics board)
<i>Relevant</i>	Will the research question make an impact, guide further research, or influence practice?

^aAdapted from Hulley et al. [4]

we do not know how to transfer a flap with a 0.8 mm luminal diameter of the vascular pedicle and (2) we do not have the required supramicrosurgical instruments. Additionally, there is the following study design barrier: (3) QoL scales may not be sensitive enough to capture the difference. Therefore, the research question may need to be reassessed based on feasibility.

Interesting

This aspect asks if your study's outcomes will interest your peers, including other surgeons, healthcare providers, and patients [4–6]. Hulley et al. [4] recommend speaking with mentors, potential funding agencies, and experts in the area to determine if your proposed research question will be of interest to a large audience.

For example, the use of suprafascial perforator flaps for postmastectomy reconstruction would be of great interest to patients. With this technique of harvesting autogenous tissue to reconstruct a breast, the patient will not worry about the potential of abdominal hernia or abdominal bulge as the abdominal muscles are not disturbed surgically. This potential lower risk of an abdominal complication would also be important to surgeons and hospital administrators in reducing corrective reoperations and ultimately, time spent, and resources used in hospital.

Novel

This criterion asks if your research will provide new information, or if it will refute, confirm, or extend previous findings. An important aspect of each research project is explaining why this project is important considering what is already known [4–6]. Also, a research question can be novel while not completely original. The novelty of your research question can be established with a literature search, which can help identify gaps and remaining questions in your research area. For example, recreating a previous study but reducing past limitations, or focusing on a new population, could also be considered novel [4].

For example, the suprafascial periumbilical perforator flaps for postmastectomy reconstruction is a novel technique for harvesting autogenous tissue for reconstruction and extends the boundary of microsurgical instrumentation and skills [9].

Ethical

While there is a certain level of risk associated with most research projects, one must ensure that the proposed research project will be amendable to approval by the local research ethics board [4–6]. If an original research question poses unacceptable risks to patients/participants, it may be possible to utilize a different study design to find the appropriate answer. Although the randomized control trial design with a narrow confidence interval would provide the highest level of evidence in a question of effectiveness, this study design would not be deemed ethical if the research question is about harm. In the example of the suprafascial periumbilical perforator flap, we cannot think of any concerns that as a novel intervention, it places the patient at a higher risk of harm. Thus, we may feel confident that a research ethics board will not question the existence of equipoise in performing a study comparing these two approaches. For a question of Harm in surgery, however, a better study design is a case-control study [10].

Relevant

Lastly, this aspect asks if the issue under consideration in a project is clinically relevant to physicians/surgeons, patients, health policy-makers, and other researchers. This aspect also focuses on whether the issue being researched is timely or applicable to the issues we face in the present or future moments [4–6]. The National Institutes of Health (NIH), in the United States of America, emphasizes the importance of a research question. By importance, we mean how the results will improve knowledge, and how results will impact clinical services, methods, and concepts [4].

For example, the use of the suprafascial peri-umbilical perforator flap will be deemed relevant and timely to the stakeholders mentioned above. An older technique for breast reconstruction such as a regional pedicled skin flap would be considered a “historical” method of reconstruction and not currently relevant.

The PICOT Format

While the FINER criteria justify a planned research project, the PICOT format helps in the development of a specific research question [6] which will help us find the answer through the literature search if we are “Users” of surgical research. If we are “Doers” of surgical research, it will help us formulate our research question before we submit it to the local ethics committee or a granting agency.

The sooner a surgeon adopts the PICOT format the earlier he/she becomes able to search the literature and find the answer to a clinical problem. For the surgeon-investigator, a clinical question posed in the PICOT format will guide research. Once a systematic review is complete and the “boundary of knowledge” is understood, the question can be formed.

The PICOT format, summarized in Table 3.2, includes the following components: (1) Patient or population, (2) intervention, (3) comparative intervention, (4) outcome, and (5) time horizon [3, 5–8]. For the surgeon-investigator, a clinical question posed in the PICOT format will guide research. The PICOT format, outlined in more detail below, demands clarification for each of its components.

Patient or Population

For the first component, *patient or population*, the type of patient or patient population we are dealing with must be clarified [3, 5–8]. Having a clear definition of the patient or patient population allows one to identify if the patient(s) referenced in an article we read are similar enough to the patient or population for whom we need the answer. For example, in the case of a patient or population with appendicitis, does our patient have acute appendicitis or chronic appendicitis? In the case of inguinal hernia, does our patient have an asymptomatic inguinal hernia or an acute strangulated one?

Intervention

In the second component of the PICOT formula, *intervention*, we would like to know, as precisely as possible, what the main intervention is [3, 5–8]. For example, if we are comparing laparoscopic versus open cholecystectomy, the laparoscopic procedure could be identified as the main intervention. In this case, we would want to know if the main intervention is a four-port laparoscopic approach or a single port laparoscopic intervention.

Comparative Intervention or Control

Using the same reasoning for the *comparative intervention*, we would like to know, again precisely, what this comparative intervention was [3, 5–8]. The comparative intervention is often either a control group or the best current standard procedure for the respective condition, sometimes called the gold standard or community standard.

Table 3.2 The PICOT format

Component	Definition
<i>Patient or population</i>	The patient population of interest
<i>Intervention</i>	What is being done for/to the patient
<i>Comparison intervention</i>	What is the intervention being compared to? Often what is currently being done or the “gold standard”
<i>Outcome</i>	What is being measured? What are you hoping to improve/see a change?
<i>Time horizon or time frame</i>	How long after the intervention do you measure the outcome?

^aTable 3.2 was formed using information from [3, 5–8]

Using the example of laparoscopic versus open cholecystectomy where the laparoscopic procedure was the main intervention, the comparative intervention would be the open cholecystectomy. For the open cholecystectomy, we may want to know how extensive the incision was; was it a Kocher incision? Was it a midline incision from the xiphoid to the suprapubic area? Or was it just limited to the epigastric area (above the umbilicus)?

Outcome

Regarding the *outcome* component of the PICOT, we want to know what was hoping to be measured, accomplished, improved, or affected [3, 5–8]. Was it a critical outcome such as mortality, was it a QoL outcome, or was it something minor such as a rash around the incision? The outcomes can be variable and may include, but not be limited to, (1) a critical outcome (mortality), (2) a QoL outcome, (3) pain, (4) hospitalization days, or (5) ability to return to work. The primary outcome of a study should be succinctly stated as it has implications in the study design, sample size, and power of the study (see Chap. 34). These issues will be discussed later in Chap. 29.

Time Horizon

The final component of the PICOT formulation addresses the *time horizon* of the study. The time horizon describes when the outcome was measured in reference to the intervention/event [4]. For example, was the outcome measured 1-month postprocedure or 1 year? A study's time horizon depends on the nature of the surgical problem we are exploring. Nobody would take us seriously if we reported our survival results of laparoscopic versus open colon cancer resection at 6 months after surgery. On the other hand, this time horizon may be quite appropriate for quality of life after appendectomy. There may be occasions, for example, when measuring pain, where an intermediate follow-up period, in addition to a long-term fol-

low-up is relevant. Even if there is no difference in survival after colon resection with open versus laparoscopic approaches, there may be a difference in pain levels or length of time hospitalized. It is important for investigators or authors of published reports or clinical trials to justify the time horizon used [11]. When all components are combined, they form a clear and searchable research question.

Below are four examples of such foreground questions with their equivalent PICOT formulation.

Example 1 In patients undergoing laparoscopic surgery for colorectal cancer, does the use of ultrasound-guided transversus abdominis plane (TAP) blocks reduce pain up to 72 h following surgery?

- P. Population: patients with colorectal cancer undergoing laparoscopic surgery
- I. Intervention: ultrasound-guided TAP blocks
- C. Comparative intervention: no TAP blocks
- O. Outcome: pain
- T. Time horizon: up to 72 h following randomization

Example 2 In patients undergoing prostatectomy for prostate cancer, is robotic-assisted prostatectomy more likely to allow patients to maintain erectile function than traditional open approach, at 1-year postsurgery?

- P. Population: patients with prostate cancer undergoing prostatectomy
- I. Intervention: robotic-assisted prostatectomy
- C. Comparative intervention: open prostatectomy
- O. Outcome: erectile function
- T. Time horizon: 1 year following randomization

Example 3 In patients 65 years or older who experience a hip fracture, does accelerated surgery (surgery within 6 h) decrease all-cause mortality compared to usual care within 30 days following surgery?

- P. Population: patients 65 years or older with a hip fracture
- I. Intervention: accelerated surgery (surgery within 6 h)

- C. Comparative intervention: standard surgical care
- O. Outcome: all-cause mortality
- T. Time horizon: 30 days from surgery

Example 4 In patients with carpal tunnel syndrome, is the endoscopic carpal tunnel release (ECTR) technique more cost-effective than the open carpal tunnel release (OCTR) technique at 1 year postsurgery?

- P. Population: patients with clinical symptoms of carpal tunnel syndrome
- I. Intervention: ECTR
- C. Comparative intervention: OCTR
- O. Outcome: cost-effectiveness (dollars per natural unit)
- T. Time horizon: 1 year

Conclusions

The development of a research question could be seen as the most important step in a research project. In this chapter, we explained two tools that can be used to ensure the formation of a clear and concise clinical research question: The FINER criteria and the PICOT format (Box 1 and Box 2). A strong research question lays the foundation of a well-designed research project. When structured properly, a research question will make your literature search more efficient through minimizing the results you encounter. Furthermore, a well-formed research question lays a strong foundation and provides direction for the surgeon investigator throughout all stages of the research project, from study design to publication.

References

- Haynes RB, Sackett DL, Guyatt GH, Tugwell P. Clinical epidemiology: how to do clinical practice research. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2006.
- Sackett D, Richardson WS, Rosenberg W, Haynes RB. How to practice and teach evidence based medicine. London, UK: Churchill Livingstone; 1997.
- Thoma A, McKnight L, McKay P, Haines T. Forming the research question. Clin Plast Surg. 2008;35(2):189–93.
- Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Conceiving the research question and developing the study plan. In: Gaertner R, editor. Designing clinical research. 4th ed. Philadelphia: Lippincott Williams & Wilkins, a Wolters Kluwer business; 2013. p. 14–22.
- Thabane L, Thomas T, Ye C, Paul J. Posing the research question: not so simple. Can J Anesth. 2009;56(1):71–9.
- Farrugia P, Petrisor BA, Farrokhyar F, Bhandari M. Research questions, hypotheses and objectives. Can J Surg. 2010;53(4):278–81.
- Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club. 1995;123(3):A12–3.
- Nollan R, Finout-Overhold E, Stephenson P. Asking compelling clinical questions. In: Melnyk BM, Fineout-Overhold E, editors. Evidence-based practice in nursing and healthcare: a guide to best practice. Philadelphia: Lippincott Williams & Wilkins; 2005.
- Koshima I, Yamamoto T, Narushima M, Mihara M, Iida T. Perforator flaps and supermicrosurgery. Clin Plast Surg. 2010;37(4):683–9.
- Thoma A, Kaur MN, Farrokhyar F, Waltho D, Levis C, Lovrics P, Goldsmith CH. Users' guide to the surgical literature: how to assess an article about harm in surgery. Can J Surg. 2016;59(5):351–7.
- Mowackett S, Karpinski M, Gallo L, Gallo M, Banfield L, Murphy J, Waltho D, Copeland A, Moltaji S, Thoma A. Reporting time horizons in randomized controlled trials in plastic surgery: a systematic review. Plastic Reconstr Surg. 2018;142(6):947e–57e.

Finding the Evidence Through Searching the Literature

4

Laura Banfield and Jo-Anne Petropoulos

Abbreviations

PICO(T)	P = population, I = intervention, C = comparator, O = outcome of interest, and T = time
BMI	Body Mass Index
MeSH	Medical Subject Heading
EBM	Evidence-Based Medicine
AI	Artificial intelligence
DOI	Digital Object Identifier

Introduction

Questions often arise that require finding evidence to inform clinical decisions. However, not all clinicians seek answers to all questions [1, 2]. Clinician reasons for not seeking answers include lack of time, questions lack urgency, importance, or are forgotten, perception that an appropriate or relevant answer does not exist, questions are patient specific and nongeneralizable, and challenges in using the resources [1–7]. “Because of the increasing volume and speed of new research, finding useful evidence efficiently remains challenging” [8]. As stated by Waltho et al. [9], “surgeons must always ensure that the care they

provide is rooted in the best available evidence” [9]. This chapter will equip surgeons with knowledge and skills to locate the best evidence to inform their practice.

Clinical Scenario

A 67-year-old female is seen in the bariatric program with morbid obesity and hypertension. Her height is 5 ft., 6 in. Her weight is 360 lbs. She was not able to reduce her weight with dieting over the last decade. She finally decided to proceed with bariatric surgery. Her bariatric surgeon recommended Roux-en-Y surgery. Her best friend, however, had a sleeve gastrectomy and was happy with the results. She is not certain which procedure to undergo. She is willing to undergo the procedure which is likely to help her lose at least 50% of her weight over the long term (>5 years). As the surgeon, you decide to review the literature to determine which of the two options will best meet the patient’s goal. However, you are unsure of how to search the literature.

Clinical Question

As outlined in Chap. 3, *Clinical Question*, a number of different questions may arise from any one scenario or clinical encounter. Remember, “a well-designed research question addresses several components of the clinical scenario,” often in

L. Banfield (✉) · J.-A. Petropoulos
Health Science Library, HSC 2B, McMaster
University, Hamilton, ON, Canada
e-mail: banfie@mcmaster.ca; petroja@mcmaster.ca

the form of PICO(T)—P as the population affected, I as the intervention, C as the comparator, comparison intervention, or standard of care, O as the outcome of interest, and T as either time or type of study [9]. For example, in this scenario, the clinical question you developed is in patients with morbid obesity and BMI >40, does Roux-en-Y, in contrast to sleeve gastrectomy, lead to more weight loss that persists at greater than 5 years?

P	morbidly obese patients with BMI >40
I	Roux-en-Y bariatric surgery
C	Sleeve gastrectomy
O	Loss of 50% weight
(T)	Greater than 5 years

Importance of the Question

A well-formulated clinical question plays a crucial role in efficiently and effectively finding the evidence. When written well, a clinical question can inform what to search (i.e., concepts) and where to search (i.e., resources).

Clinical questions can be further broken down into background and foreground questions. Background questions typically begin with who, what, when, where, and how (e.g., the pathophysiology of a disease and explanation of a surgical procedure). These questions can be answered in textbooks or reference books. Foreground questions seek to address a specific clinical problem (e.g., causal or risk factors, prognosis, treatment effectiveness as it pertains to a specific patient or patient population). The answers to these questions can be found through effective use of resources supporting evidence-based medicine such as those within the Pyramid of EBM Resources or 6S Pyramid (see Choosing Appropriate Resources) [8, 10]. Although some of the resources found within the pyramids can be used to find background information, the focus of this chapter will be on how to use them to find evidence to support clinical decisions and research.

We acknowledge that even a well-written clinical question may not have an answer. Potential reasons could include: “no feasible study design

or measurement tools exist that investigators could use to resolve an issue,” or “no one has conducted and published the necessary study” [8]. In these instances, you will need to reevaluate your question (see Chap. 3).

Developing a Search Strategy

This section describes the core principles for developing a search strategy. As mentioned previously, a key step to an efficient search is formulating a well-written question. This is followed by identifying search concepts and their related search terms and then combining them using the Boolean operators; “herein lies the art to efficient searching” [10].

Identifying Concepts

Begin with identifying concepts from your question. This is important for developing an effective search strategy. Concepts can be more easily identified once the clinical question has been written using the format of PICO(T). Each of the components of PICO(T) become your concepts (Table 4.1).

Identifying Search Terms

After you have identified your concepts, think of alternative terms, known as keywords or text words, for each concept. Keywords are individual words or phrases that may appear in the title, abstract, author-supplied keywords, or the body of the text depending on the nature of the resources being searched. Consider including synonyms or related terms relevant to each concept. Be sure to incorporate variations of the root words as part of your list of keywords (e.g., anesthesia, anaesthesia, anesthetic, anaesthetic). An efficient way to include these variations is by adding an asterisk (*), known as a truncation symbol or a wildcard (e.g., anesthe*, anaesthes*) (Table 4.2). Be mindful of which words you apply the asterisk to as it may result in cap-

Table 4.1 Clinical question broken into PICO(T) concepts

<u>Concept A</u>	<u>Concept B</u>	<u>Concept C</u>	<u>Concept D</u>	<u>Concept E</u>
P	I	C	O	T
Morbidly Obese Patient (BMI > 40kg/m ²)	Roux-en-Y Bariatric Surgery	Sleeve Gastrectomy	Weight Loss	> 5 years

Table 4.2 Search concepts and search terms

Concept A	Concept B	Concept C	Concept D	Concept E
P	I	C	O	T
Morbidly Obese Patient (BMI > 40kg/m ²)	Roux-en-Y Bariatric Surgery	Sleeve Gastrectomy	Weight Loss	> 5 years
Obese	Gastric Bypass	Gastric Sleeve Surgery	Weight Reduc*	
Over weight	Gastroileal Bypass		Decrease* Weight	

turing words you may not have intended (e.g., stud* will retrieve stud or studs or study or studies or student or students). Some resources also allow you to include a form of internal truncation, sometimes referred to as wildcards. The purpose of internal truncation is to allow for variation in spelling (e.g., an?esthesia)—a question mark (?) or a number sign (#) may be used depending on the database.

Advanced Search Tip: Generating Additional Search Terms

To generate more search terms, take a known article and look at the title and abstract for words related to your concepts and the author-supplied keywords at the bottom of the abstract.

Another strategy is to look up the article in a database (e.g., PubMed) to identify relevant controlled vocabulary.

Controlled Vocabulary

Some resources offer alternatives to keyword only searches; this is referred to as searching with controlled vocabulary (e.g., subject headings, descriptors). Controlled vocabulary are predefined terms (labels) set by the databases to facilitate finding information. These terms are used to describe the contents of an article and are helpful when authors use different terms to describe a particular concept. For example, an author might refer to one of the following terms to describe our concept of gastric bypass surgery: gastric bypass or gastroileal bypass or gastrojejunostomy or Roux-en-Y gastric bypass (Table 4.3). Controlled vocabulary recognizes that these terms are all referring to the same concept. Instead of searching on every single term, you can search on the controlled vocabulary term to retrieve articles that use one of the four terms in our example. It is important to note that

Table 4.3 Controlled vocabulary term for gastric bypass in PubMed (MeSH term) and related terms covered by gastric bypass

MeSH	Gastric bypass
Coverage	Bypass, Gastric Roux-en-Y Gastric Bypass Bypass, Roux-en-Y Gastric Gastric Bypass, Roux-en-Y Roux-en-Y Gastric Bypass Greenville Gastric Bypass Gastric Bypass, Greenville Gastroileal Bypass Bypass, Gastroileal Gastrojejunostomy Gastrojejunostomies

a comprehensive search should include a combination of keywords and controlled vocabulary terms.

Advanced Search Tip: Controlled Vocabulary
Not all databases include controlled vocabulary. For those that do, there is no single universal controlled vocabulary. Rather, there are many sets of controlled vocabulary; some are unique to the database (e.g., Emtree for Embase) and some are used by more than one database (e.g., MeSH [Medical Subject Heading] is used by MEDLINE®, PubMed, and Cochrane Database of Systematic Reviews).

Boolean Operators

You can now begin to create a search strategy. Boolean operators can be used to logically organize concepts and keywords. The primary Boolean operators are AND, OR, and NOT (Fig. 4.1). In general, the AND operator is used to combine search concepts (Table 4.4). It allows you to retrieve evidence containing two or more concepts. For example, a search using Obesity AND Gastric Bypass will retrieve evidence con-

taining both concepts. The OR operator is most often applied when combining related keywords within a search concept. Using the OR operator will retrieve evidence containing any of the keywords in isolation or where they may show up together (Table 4.4). For example, a search using Surgery OR Gastric Bypass will retrieve evidence containing either or both keywords related to the same concept. In contrast, the NOT operator will explicitly exclude articles containing those terms from the result set. For example, a search using NOT animals will eliminate all articles with a reference to animals. The NOT operator should be used with caution as it could eliminate relevant articles that incidentally mention the term you are excluding. Using the same example, NOT Animals will also eliminate articles that reference both animals and humans (i.e., studies that incorporate human and animal models).

Using either the AND or the NOT operators will narrow the scope of the search and decrease the number of articles retrieved (see Limits section); also referred to as specificity. On the other hand, the OR operator will broaden the scope of the search and increase the number of articles retrieved, also referred to as sensitivity.

Advanced Search Tip: Applying Boolean Operators
Typically, search concepts are combined using the AND operator. However, there are circumstances in which they can be combined using the OR operator (e.g., Intervention (Roux-en-Y bariatric surgery) OR Comparator (Sleeve gastrectomy). This is particularly applicable to surgical questions.
Within each concept, the alternative terms and controlled vocabulary are usually combined using the OR operator.

Fig. 4.1 Boolean operators. *Note* Shading indicates the result set when concepts are combined using AND OR NOT

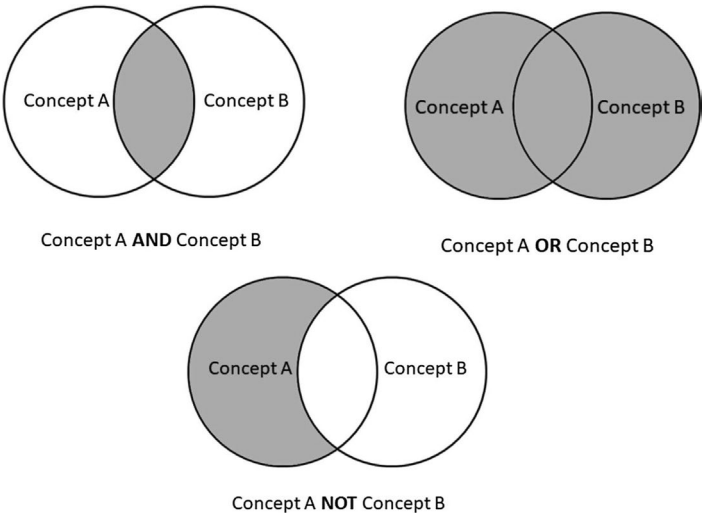


Table 4.4 Applying Boolean operators to search concepts and search terms

Concept A	Concept B	Concept C	Concept D	Concept E
P	I	C	O	T
AND	AND	AND	AND	
Morbidly Obese Patient (BMI > 40kg/m ²)	Roux-en-Y Bariatric Surgery	Sleeve Gastrectomy	Weight Loss	> 5 years
Obes*	Gastric Bypass	Gastric Sleeve Surgery	Weight Reduc*	
OR	OR		OR	
Over weight	Gastroileal Bypass		Decrease* Weight	
OR	OR			
Overweight	Greenville Gastric Bypass			
	OR			
	Gastrojejunostomies			

Creating a Search Strategy

For the purposes of this section, we have identified two approaches to searching: simple and complex. These approaches can be applied to the sources of evidence identified in the

Pyramid of EBM Resources and 6S Pyramid (see Choosing Appropriate Resources). The nature of the resource will dictate the approach to searching. See Table 4.5 for a list of resources in which these search strategies can be applied.

Artificial Intelligence (AI)

Artificial intelligence (AI) is technology that enables computers and machines to simulate human intelligence, such as problem-solving or learning. AI includes the subfield of machine learning, which includes the subfield of deep learning. AI tools may be used to identify relevant journal articles, summarize an article's key findings, generate a response to a clinical query, or rank records in information resources [11].

The use of AI tools or the AI application(s) in some of the resources listed in Table 4.5 to identify relevant literature or generate a response should be done with caution. Some limitations of AI include the exclusion of grey literature and articles without digital object identifiers (DOIs), and the inability to capture nuanced meanings and replicate a search strategy.

search as above using the search function or to use the PICO tool. In either instance, you will use the limits on the result screen to look at guidelines specifically. Known associations and organizations (professional, health) are good starting points for guidelines and may have them listed in the documents or publications sections of their websites. Alternately, guidelines may be published in journals and thus it may be more effective to search the article databases and include guideline as a concept.

When searching with one or two concepts, you scan or scroll the summary to obtain more detailed information incorporating the other aspects of your PICO(T). For many resources within these levels, you may apply a similarly simple search. However, there are some exceptions, specifically, searching for guidelines.

Simple Search Strategy (i.e., Basic)

A simple search strategy can be applied to any resource using one or two concepts, one term per concept (Fig. 4.2). However, it is best applied at the top levels in both pyramids (e.g., specifically DynaMed Plus, UptoDate®).

Using the concepts from our PICO(T) (Table 4.1), the search strategy could be any one of the following but not limited to:

1. Obesity
2. Roux-en-Y bariatric surgery
3. Sleeve gastrectomy
4. Obese AND sleeve gastrectomy

Advanced Search Tip: Searching for Clinical Practice Guidelines

Searching for guidelines can be achieved through meta-search tools such as TRIP, association or organization websites, or through the article databases. The TRIP database allows you to construct a simple

Complex Search Strategy (i.e., Advanced)

A complex search strategy can be applied to many of the highlighted resources related to the Pyramid of EBM Resources and 6S Pyramid and to other resources outside of it. This type of strategy involves two or more concepts and may also involve more than one term, including controlled vocabulary, per concept. It is best applied when searching for syntheses and studies (i.e., non-preappraised or syntheses and studies levels).

Again, using the concepts from our PICO(T), the search strategy could be any one of the following but not limited to:

1. (Obes* OR overweight OR over-weight) AND (sleeve gastrectomy OR gastric band* OR lap-band)
2. (sleeve gastrectomy OR gastric band* OR lap-band) AND (Roux-en-Y bariatric surgery OR RYGB)

Table 4.5 List of resources for simple search and complex search

Resource	Website	Simple search	Complex search	AI assistance (tools may be clearly AI or involve a “pinch” of AI)
ACCESSSS	https://www.accessss.org/	✓		
BMJ Best Practice ^a	https://bestpractice.bmj.com/	✓		
CINAHL Database ^a	https://health.ebsco.com/products/the-cinahl-database/allied-health-nursing	✓	✓	
Cochrane Central Register of Controlled Trials (CENTRAL) ^a	http://www.cochranelibrary.com/about/central-landing-page.html	✓	✓	
Cochrane Database of Systematic Reviews (CDSR) ^a	http://www.cochranelibrary.com/cochrane-database-of-systematic-reviews/	✓	✓	
DynaMed Plus ^a	http://www.dynamed.com/	✓		✓
Embase ^a	https://www.elsevier.com/solutions/embase-biomedical-research	✓	✓	
Emcare ^a	https://www.wolterskluwer.com/en/solutions/ovid/ovid-emcare-14007	✓	✓	
Epistemonikos	https://www.epistemonikos.org/	✓		✓
Essential Evidence Plus ^a	https://www.essentialevidenceplus.com/	✓		
LILACS	http://lilacs.bvsalud.org/en/	✓	✓	
MEDLINE® ^a	https://www.nlm.nih.gov/bsd/medline.html https://www.nlm.nih.gov/medline/	✓	✓	
OrthoEvidence™ ^a	https://myorthoEvidence.com/	✓		
PEPID ^a	https://www.pepidconnect.com/	✓		
PsycINFO® ^a	http://www.apa.org/pubs/databases/psycinfo/		✓	
PubMed	https://www.ncbi.nlm.nih.gov/pubmed/	✓	✓	✓
Scopus ^a	https://www.elsevier.com/solutions/scopus		✓	✓
TRIP ^b	https://www.tripdatabase.com/	✓		✓
UpToDate® ^a	https://www.uptodate.com/contents/search	✓		
CMA CPG Infobase ^b	https://www.cma.ca/En/Pages/clinical-practice-guidelines.aspx	✓		
International Guideline Library	https://www.g-i-n.net/library/international-guidelines-library/international-guidelines-library	✓	✓	
WHO Guidelines	http://www.who.int/publications/guidelines/en/	✓		
NICE Guidance	https://www.nice.org.uk/	✓		

Note: Access to fee-based resources may be provided through your professional, academic, clinical, and hospital affiliations

^afee-based

^bpremium or extra features with additional fee

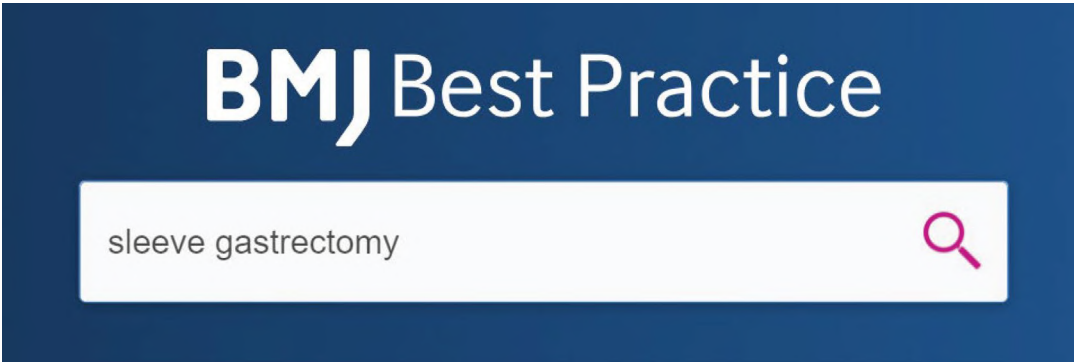


Fig. 4.2 Applying a simple search to BMJ best practice

Table 4.6 Sample of a complex search using OVID MEDLINE®

Line	Search terms
1	Obesity/
2	obes*.mp.
3	over weight.mp.
4	overweight.mp.
5	1 or 2 or 3 or 4
6	sleeve gastrectom*.mp.
7	Gastric Bypass/
8	roux-en-y bariatric surgery.mp.
9	7 or 8
10	5 and 6 and 9

Note The / indicates use of controlled vocabulary (MeSH) and.mp. indicates use of keywords. Line 5 illustrates the combining of terms used to describe the population. Line 6 illustrates the comparison, while Line 9 illustrates the combining of terms used to describe the intervention. Line 10 is the combining of all 3 concepts by applying AND. This search could be built upon through the addition of more related terms and concepts

3. (bariatric surg*) AND (weight loss or weight reduc*) (Table 4.6)

Advanced Search Tip: Selecting Search Concepts

In PICO(T) a comparator (C) or time period/type of study (T) may not always be part of your search strategy. These can be very helpful when reviewing your search results. The more concepts you search on, the more specific your search is, and the narrower your results become.

Limits and Search Filters

Most resources enable you to limit or narrow your search results. The limit options available vary across the resources. Examples of limits include publication date, language, age, and publication type. It is best to apply these limits one at a time, and at the end of your search. This will allow you to see the impact that each of the limits will have on narrowing your results. In some circumstances, it may be necessary to revisit the limits you have chosen based on your preliminary results. This is easier to do if the limits have been applied one at a time.

Advanced Search Tip: Limits as a Search Concept

A helpful strategy is to treat a limit as a search concept. This will increase the sensitivity of the search by allowing you to use more terms to describe it (e.g., Randomized Controlled Trial* or RCT* or blind* or conceal*...). This advanced approach can also involve using controlled vocabulary in addition to keywords.

Filters, also known as hedges, are combinations of predefined search terms used to retrieve results on a particular topic (e.g., cancer), with a particular study design (e.g., randomized controlled trials or cohort), or using a particular study methodology (e.g., prognosis or therapy).

Filters are applied to improve the quality of the result set by reducing the number of irrelevant results and increasing the number of relevant results [12–17]. These filters are added to the search terms being used for the content.

PubMed, Ovid MEDLINE®, Embase, and PsycINFO®, and EBSCO CINAHL and MEDLINE®, have all adopted the search filters (clinical hedges) developed by Health Information Research Unit (HiRU) at McMaster University [2, 18]. The HiRU hedges, known as Clinical Queries, are found under the Additional Limits in Ovid® and are found within the Edit feature in the search history and Refine Results section of EBSCO CINAHL. In PubMed, the link to Clinical Queries is located in the PubMed Tools section on the homepage. The common clinical queries filters cover the following categories: therapy, diagnosis, etiology, prognosis, and clinical prediction guides. Additional HiRU hedges include reviews, economics, qualitative, and costs [18].

Advanced Search Tip: Additional Search Filters

Not all filters are created equally. It should be noted that not all filters have been developed using a rigorous methodology nor have all been validated. As a result, you should look critically at those available before selecting one that best suits your needs.

HiRU Search Filters

https://hiru.mcmaster.ca/hiru/HIRU_Hedges_home.aspx

Scottish Intercollegiate Guidelines Network (SIGN)

<https://www.sign.ac.uk/what-we-do/methodology/search-filters/>

InterTASC Information Specialists' Sub-Group (ISSG) Search Filter Resource

<https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/>

CADTH Search Filters Database

<https://searchfilters.cadth.ca/>

Choosing Appropriate Resources

“As a surgeon using EBM [Evidence-Based Surgery] at point of care, it is critical for you to ensure you are taking advantage of the best resources for making clinical and surgical decisions” [9]. This section will provide an overview of selected resources and the differences between them. Strengths and weaknesses of the resources will not be discussed as such assessments are partially based on personal preferences, availability, and access. However, it is worth noting that many of these may not have comprehensive coverage of surgical content.

There are many ways to describe and organize sources of evidence. Two of the more common ways are Pyramid of EBM Resources and 6S hierarchy of evidence-based resources Pyramid [8, 19].

Pyramid of EBM Resources

The Pyramid of EBM Resources divides resources into three categories: summaries and guidelines, preappraised research, and non-preappraised research.

Summaries and Guidelines

Summaries and guidelines are rarely limited to a specific clinical question. Instead, they are topic or condition driven, often integrating background and foreground information to inform clinical decision-making [8]. These resources rate the strength of the recommendations based on either or both internally defined criteria and externally defined criteria (e.g., GRADE). Examples of these resources include UpToDate®, DynaMed Plus, BMJ Best Practice, and clinical practice guidelines.

Preappraised Research

Preappraised research represents a single page synopsis of a systematic review or single study. The synopsis includes an evaluation of the quality of the methods, clinical relevance, and expert

commentary [8]. Examples include ACP Journal Club and OrthoEvidence™.

Non-preappraised Research

Non-preappraised research refers to primary studies with no evaluation [8]. Examples include articles found in PubMed, Embase, MEDLINE®, and Cochrane Central Register of Controlled Trials (CENTRAL).

6S Pyramid

The 6S Pyramid takes a slightly different approach to organizing the sources of evidence, yet the similarities between the two are evident (Fig. 4.3). Both pyramids place studies at the bottom and move up towards more clinically driven resources. The key differences include the addition of a Systems layer at the top and the apparent subdivision within the preappraised research and non-preappraised research categories in the 6S Pyramid.

Systems

Systems represent an ideal situation in which the best available evidence informing clinical decision making is integrated with the information unique to each patient through the electronic medical record [19]. To our knowledge, there are few in existence. We recognize their role in and

potential to contribute to evidence-based patient care, but we will not explore them further.

Summaries

As with the Summaries and Guidelines level from the Pyramid of EBM Resources, this level provides “a comprehensive view of the body of evidence at a topic level” [8]. Examples of these resources include UptoDate®, DynaMed Plus, BMJ Best Practice, Essential Evidence Plus, Micromedex®, PEPID, and clinical practice guidelines.

Synopses of Syntheses

Synopses of syntheses are typically one-page structured summaries that provide an appraisal of a systematic review. They are explicit in their assessment of the quality and will often note newsworthiness and relevance to clinical practice and research. Examples of resources relevant to surgery in this category include OrthoEvidence™, ACP Journal Club, and Cochrane evidence summaries.

Syntheses

Commonly referred to as a systematic review, a synthesis is a comprehensive summary of all the evidence surrounding a particular research question. Examples of where systematic reviews can be found include individual journals and through the Cochrane Database of Systematic Reviews, PubMed, MEDLINE®, and Embase.

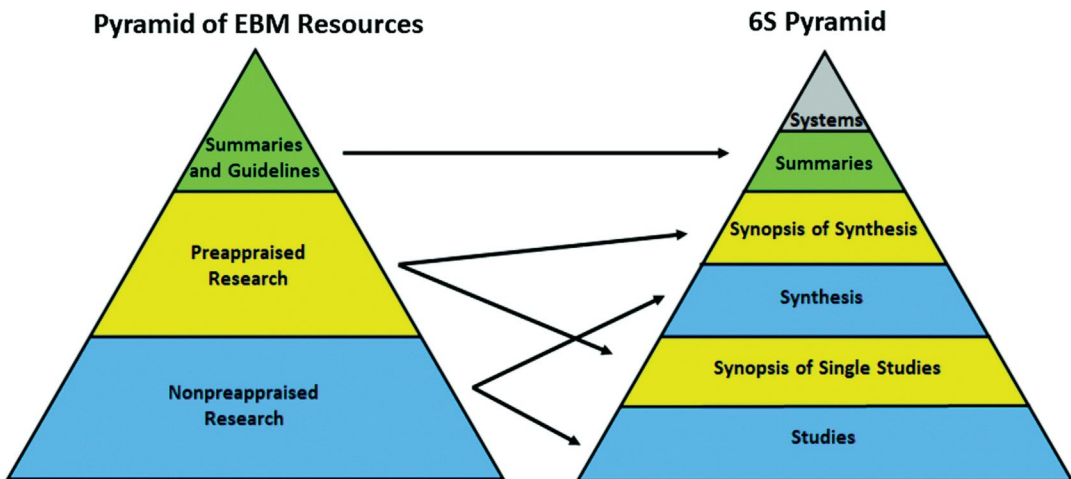


Fig. 4.3 Relationship of the Pyramid of EBM Resources to the 6S Pyramid (Used with permission)

Synopses of Single Studies

Synopses of original studies are typically one-page structured summaries that provide an appraisal of single studies. Similar to synopses of synthesis, they are explicit in their assessment of the quality of the study and will often note newsworthiness and relevance to clinical practice and research. Examples of resources include OrthoEvidence™, ACP Journal Club, and BMJ Evidence-based Medicine.

Studies

Studies are single articles reporting on original research. Unlike the resources above, the responsibility for critically appraising or assessing the quality of the evidence falls to the reader. Examples of where studies can be found include individual journals and through databases including PubMed, MEDLINE®, Embase, and Cochrane Register of Controlled Trials (CENTRAL).

Other Resources

After reviewing the resources described, it is common to question the necessity of searching multiple resources. If your purpose in searching is emerging out of a clinical situation to which you need a more immediate answer, the preappraised and summaries resources are likely to be your best starting point (Fig. 4.3). However, if you are undertaking a systematic review, clinical practice guideline, or other form of knowledge synthesis, you are more likely to be working with the non-preappraised resources (synthesis and studies). Before beginning your search consider why you are searching, what resources you have access to, and how much time you have.

Federated Search Tools

These are resources that search across multiple sources of evidence simultaneously are referred to as federated or meta-search tools. The search results from a number of resources are dis-

played together and may be broken up into categories depending on the organization of the tool (Fig. 4.4). Examples include TRIP database, Epistemonikos, ACCESSSS, and SumSearch.

Google and Google Scholar

A few words on Google and Google Scholar: both are powerful, convenient, and easy to use search tools. Google can be very helpful in identifying information about rare conditions, procedures, and emerging trends and outbreaks, as well as in identifying reports and policy statements from governmental and nongovernmental organizations and agencies, research institutes and think tanks. Google Scholar is helpful in identifying research literature across disciplines and articles in press.

However, they should be used with caution. Google searches across a wide range of information and web-based communication, not all of which is scientific, evidence-driven or without some form of bias (e.g., personal opinion, commercially driven). It is not easy to determine the good from the bad without employing strong skills in critical appraisal. The lack of transparency in the search algorithm and the uncertainty in how current it is are some concerns with Google Scholar. While both are easy to use, it can be rather time-consuming sifting through the results looking for evidence that can be applied to your question, which outweighs the ease of searching.

Access to Resources

Depending on your setting and your academic, clinical, and professional affiliations, you may have access to fee-based resources (e.g., Ovid MEDLINE®, PEPID) in addition to the “free” resources such as TRIP, PubMed, and Epistemonikos (Table 4.5). At the time of writing, some of “free” resources were offering premium and extra features through an additional fee (e.g., TRIP PRO provides images, videos, and greater search refinement).

obesity AND sleeve gastrectomy

3,453 results Pro users have access to +385 Systematic Reviews

Filter Results

FROM

TO

Apply

All Secondary Evidence

Systematic Reviews 28

Evidence Based Synopses 14

Guidelines

- Aus. & NZ 6 Pro
- Canada 5 Pro
- UK 5 Pro
- USA 29 Pro
- Europe 12 Pro
- Other 11 Pro

Regulatory Guidance 5

Key primary research 38

Clinical Q&A 1

Controlled Trials 304

Primary research 2798

Ongoing Systematic Reviews 108

Ongoing Clinical Trials 714 Pro

Patient Information Leaflets 6

Blogs 4

eTextbooks 57

Order By

Quality

Show Snippets

Results page key

Export Articles

1

Endoscopic sleeve gastroplasty for obesity

Regulatory Guidance

2024

National Institute for Health and Clinical Excellence - Interventional Procedures

Broken Link?

2

Single-anastomosis duodeno-ileal bypass with sleeve gastrectomy for treating morbid obesity

Regulatory Guidance

2016

National Institute for Health and Clinical Excellence - Interventional Procedures

Broken Link?

3

Short-Term Changes in Weight, Body Composition, and Metabolic Biomarkers After Laparoscopic Sleeve Gastrectomy in Patients with Obesity: A Comparative Prospective Study.

Primary research

2024

Obesity Surgery

Broken Link?

4

Side-to-side magnetic duodeno-ileostomy in adults with severe obesity with or without type 2 diabetes: early outcomes with prior or concurrent sleeve gastrectomy.

Primary research

2024

Surgery for Obesity and Related Diseases

Broken Link?

5

Sleeve gastrectomy for the treatment of adolescent obesity in children aged 13 and under: a retrospective study.

Primary research

2024

Surgery for Obesity and Related Diseases

Fig. 4.4 Result set from TRIP database (Used with permission)

Keeping Current

Staying current with the latest evidence keeps you informed of what others in your field are doing, assists in contextualizing your work and helps with identifying research and knowledge gaps. Many of the previously mentioned resources will allow you to stay current with the

latest evidence. Depending on the resource, you can create table of contents alerts, search alerts, and alerts informed by predefined categories (e.g., discipline, topic, profession) (Fig. 4.5). These alerts can be customized to frequency and timing of delivery, number of results, and format of the alert (e.g., email, RSS, citation management software compatible).

Personalized Alert Settings

Relevance cut-off score * ⓘ

4

Disciplines *

Click here to select

Newsworthiness cut-off score * ⓘ

4

Populations *

Click here to select

Email alert frequency *

Weekly

Exclude COVID Alerts * ⓘ

☐ Yes

☒ No

Save changes

Select your preferences above to calculate the average number of articles per month you should expect to receive.

To receive alerts via RSS, copy this address into your RSS Reader program:

https://www.accessss.org/Rss/UserFeed/54fd7062-3c90-452c-abfa-43d03fce88bb

Disciplines *

Click here to select

☐ Select All

☐ --Dermatology

☐ --Endocrine

☐ --Gastroenterology

☐ --Genetics

☐ --Geriatrics

☐ --Hematology

☐ --Hemostasis and Thrombosis

☐ --Infectious Disease

Populations *

Click here to select

☐ Select All

☐ Neonates

☐ Pediatrics

☐ Adolescents

☐ Adults

☐ Geriatrics

Fig. 4.5 Alert settings from ACCESSSS (Used with permission)

The screenshot shows the 'Personalized Alert Settings' form. It includes fields for 'Relevance cut-off score' (set to 4), 'Newsworthiness cut-off score' (set to 4), 'Email alert frequency' (set to Weekly), and 'Exclude COVID Alerts' (set to No). A green 'Save changes' button is at the bottom. Below the button is a message: 'Select your preferences above to calculate the average number of articles per month you should expect to receive.' At the bottom, there is an RSS feed link: 'https://www.accessss.org/Rss/UserFeed/54fd7062-3c90-452c-abfa-43d03fce88bb'. Two callout boxes are shown: one for 'Disciplines' and one for 'Populations'. The 'Disciplines' callout shows a list of medical specialties with checkboxes: Select All, --Dermatology, --Endocrine, --Gastroenterology, --Genetics, --Geriatrics, --Hematology, --Hemostasis and Thrombosis, and --Infectious Disease. The 'Populations' callout shows a list of population groups with checkboxes: Select All, Neonates, Pediatrics, Adolescents, Adults, and Geriatrics.

Fig. 4.5 (continued)

Resolution of the Clinical Scenario

You are now ready to resolve the clinical scenario by applying what you have learned in this chapter. The approach you take will be informed by your purpose in searching, what you have access to, and the amount of time you have. These factors together with either of the pyramids discussed will help to identify which resources are more appropriate to search given your circumstances.

If the purpose is to inform patient care, we recommend that you start with the resources at the top of the pyramid and work downwards until you find a suitable answer. Alternatively, if the purpose is to engage in research, we recommend that you start with the resources at the bottom of the pyramid and take into account existing synthesis.

Applying a simple search (e.g., Roux-en-Y, sleeve gastrectomy) in DynaMed Plus leads to the evidence summary entitled Bariatric Surgery [20]. Resources at the summaries level, such as DynaMed Plus, tend to be topic- or concept-driven. They contain both background and fore-

ground information and require you to read within the entry to find the evidence.

Building upon the complex search strategy from Ovid Medline (Table 4.6) and adding limits for age and humans, a 2017 article comparing sleeve gastrectomy to Roux-en-Y in older adults by Casillas et al. is found [21]. As this is a prospective cohort study, you will need to apply the critical appraisal skills discussed in Chap. 17.

Conclusion

Surgeons encounter questions that require them to search for evidence to inform a clinical decision or research process. There are many sources of information available and different search strategies that can be applied. This chapter has covered the core principles for developing a search strategy from identifying concepts, search terms, and controlled vocabulary (if applicable), to combining concepts and terms using Boolean operators, and selecting resources. However, no single resource will be sufficient to address all of

your information needs and it may be necessary to search multiple sources. Federated search tools and AI applications can be invaluable especially when time is a factor.

Medical and health sciences library employees are also a key resource in assisting you with developing your searching skills and with searching for complex topics or difficult to answer questions. We encourage you to reach out, they will be happy to help you.

Paraphrasing Glasziou, Burls, and Gilbert, the ability to effectively search the literature is as essential as the stethoscope to a clinician's practice or the scalpel to a surgeon [22].

References

1. Del Fiore G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care: a systematic review. *JAMA Intern Med.* 2014;174(5):710–8.
2. Brassil E, Gunn B, Shenoy AM, Blanchard R. Unanswered clinical questions: a survey of specialists and primary care providers. *J Med Libr Assoc.* 2017;105(1):4–11.
3. Cook DA, Sorensen KJ, Linderbaum JA, Pencille LJ, Rhodes DJ. Information needs of generalists and specialists using online best-practice algorithms to answer clinical questions. *J Am Med Inform Assoc.* 2017;24(4):754–61.
4. Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME. Answering physicians' clinical questions: obstacles and potential solutions. *J Am Med Inform Assoc.* 2005;12(2):217–24.
5. Ely JW, Osheroff JA, Ebell MH, Chambliss ML, Vinson DC, Stevermer JJ, Pifer EA. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *BMJ.* 2002;324(7339):710.
6. Green ML, Ruff TR. Why do residents fail to answer their clinical questions? A qualitative study of barriers to practicing evidence-based medicine. *Acad Med.* 2005;80(2):176–82.
7. Bennett NL, Casebeer LL, Zheng S, Kristofco R. Information-seeking behaviors and reflective practice. *J Contin Educ Heal Prof.* 2006;26(2):120–7.
8. Agoritsas T, Vandvik PO, Neumann I, Rochwerg B, Jaeschke R, Hayward R, Guyatt G, McKibbin A. In: Guyatt GH, Rennie D, Meade MO, Cook DJ, editors. *Finding the best evidence. Users' guides to the medical literature: a manual for evidence-based clinical practice.* 3rd ed. New York: McGraw-Hill Education; 2015. p. 29–49.
9. Waltho D, Kaur MN, Haynes RB, Farrokhhyar F, Thoma A. Users' guide to the surgical literature: how to perform a high-quality literature search. *Can J Surg.* 2015;58(5):349–58.
10. Herbert R, Jamtvedt G, Hagen KB, Mead J. *Practical evidence-based physiotherapy.* 2nd ed. Edinburgh: Elsevier/Churchill Livingstone; 2011.
11. Kiester L, Turp C. Artificial intelligence behind the scenes: PubMed's Best Match algorithm. *J Med Libr Assoc.* 2022;110(1):15–22.
12. McKibbin KA, Wilczynski NL, Haynes RB, Hedges Team. Retrieving randomized controlled trials from medline: a comparison of 38 published search filters. *Health Inf Libr J.* 2009;26(3):187–202.
13. McKibbin KA, Lokker C, Wilczynski NL, Haynes RB, Ciliska D, Dobbins M, Davis DA, Straus SE. Search filters can find some but not all knowledge translation articles in MEDLINE: an analytic survey. *J Clin Epidemiol.* 2012;65(6):651–9.
14. Lokker C, Haynes RB, Wilczynski NL, McKibbin KA, Walter SD. Retrieval of diagnostic and treatment studies for clinical use through PubMed and PubMed's Clinical Queries filters. *J Am Med Inform Assoc.* 2011;18(5):652–9.
15. Shariff SZ, Sontrop JM, Haynes RB, Iansavichus AV, McKibbin KA, Wilczynski NL, Weir MA, Speechley MR, Thind A, Garg AX. Impact of PubMed search filters on the retrieval of evidence by physicians. *CMAJ.* 2012;184(3):E184–90.
16. EBSCO Help. CINAHL Clinical Queries [Internet]. [cited 2018 June 28]. Available from: https://help.ebsco.com/interfaces/CINAHL_MEDLINE_Databases/CINAHL/CINAHL_Clinical_Queries. Accessed 28 June 2018.
17. Scottish Intercollegiate Guidelines Network (SIGN). Search filters [Internet]. [cited 2018 June 25]. Available from: <http://sign.ac.uk/search-filters.html>.
18. Health Information Research Unit, McMaster University. Hedges [Internet]. [cited 2018 May 11]. Available from: https://hiru.mcmaster.ca/hiru/hiru_hedges_home.aspx.
19. DiCenso A, Bayley L, Haynes RB, ACP Journal Club. Editorial: accessing preappraised evidence: fine-tuning the 5S model into a 6S model. *Ann Intern Med.* 2009;151(6):JC3–2–3.
20. DynaMed Plus [Internet]. Ipswich (MA): EBSCO Information Services. 1995. Record No. 483434, Bariatric surgery; [updated 2018 July 18, cited July 20, 2018]; [about 62 screens]. Available from <http://www.dynamed.com/login.aspx?direct=true&site=DynaMed&id=483434>. Registration and login required.
21. Casillas RA, Kim B, Fischer H, Zelada Getty JL, Um SS, Coleman KJ. Comparative effectiveness of sleeve gastrectomy versus Roux-en-Y gastric bypass for weight loss and safety outcomes in older adults. *Surg Obes Relat Dis.* 2017;13(9):1476–83.
22. Glasziou P, Burls A, Gilbert R. Evidence based medicine and the medical curriculum. *BMJ.* 2008;337:a1253.

Hierarchy of Evidence in Surgical Research

5

Gina Del Fabbro, Sofia Bzovsky, Achilles Thoma,
and Sheila Sprague

Abbreviations

LOE	Levels of Evidence
JBJS	<i>Journal of Bone and Joint Surgery</i>
RCT	Randomized Controlled Trial
SR	Systematic Review
MA	Meta-Analysis
CSM	Computer Simulation Models
QALYs	Quality-Adjusted Life Years

Introduction

When answering a clinical question and applying the principles of evidence-based surgery, critical evaluation of the available evidence is of vital importance. The first step in this process is to determine where the study falls on the hierarchy of evidence. Multiple versions of the hierarchy of evidence tables have been developed since the advent of evidence-based medicine and evidence-

based surgery. The levels of evidence (LOE) were first introduced by the Canadian Task Force on Periodic Health Examination founded in 1976 at the request of the Conference of Deputy Ministers of Health across Canada [1]. Their proposal of an evidence rating system was later improved upon by David Sackett, who is considered the father of evidence-based medicine and developed five classifications for the LOE [2]. Since then, a more rigorous and complex LOE has been developed by the Centre for Evidence-Based Medicine in Oxford, United Kingdom, which ranks articles according to the study design used to answer the primary research question (see Appendix) [3]. Multiple journals and groups have adapted this table or a variation of it for their specialities. The *Journal of Bone and Joint Surgery (JBJS)* first published their version of the LOE in 2003, basing their system off of an earlier version of the Center for Evidence-Based Medicine's LOE scheme [4]. *JBJS* has since pub-

G. Del Fabbro · S. Bzovsky
Division of Orthopaedic Surgery, Department of
Surgery, McMaster University, Hamilton, ON, Canada
e-mail: delfabbg@mcmaster.ca;
bzovskys@mcmaster.ca

A. Thoma
Division of Plastic Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: athoma@mcmaster.ca

S. Sprague (✉)
Division of Orthopaedic Surgery, Department of
Surgery & Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: sprags@mcmaster.ca

lished an updated version of their LOE, which has been found to be responsive to surgical publications and highly relevant, as it addresses multiple areas of research [5]. Although we chose to focus on the *JBJS* scheme, this information is applicable to all surgical specialties. In this chapter, we review the hierarchy of evidence for therapeutic research, prognostic research, diagnostic test evaluation, and economic research, as described in *JBJS* [6].

Levels of Evidence for Therapeutic Research

Therapeutic research evaluates the effect of different treatments on one or more outcomes of interest. Simply stated, therapeutic research addresses questions, such as *does this treatment help* and *what are the harms?* [4]. Figure 5.1 shows the hierarchy of evidence for therapeutic research [4]. High-quality randomized controlled trials (RCTs), as well as systematic reviews (SR) and meta-analyses (MA) of well-conducted RCTs, are considered the highest LOE for therapeutic research. Randomized controlled trials, in which patients are randomly allocated to an intervention group or a comparison group, are considered the highest LOE because they offer the best protection against bias (Fig. 5.2a). Randomization is an effective method of reducing bias as it bal-

ances the treatment groups with respect to both known and unknown prognostic factors [7, 8]. Randomized controlled trials and subtypes are discussed more in Chaps. 11, 13, 14, and 15. Meta-analyses and SRs of high-quality RCTs also represent Level I evidence. An SR is an investigation of relevant literature related to a specific research question. SRs are conducted according to a protocol that details the search strategy, eligibility criteria for included studies, as well as data abstraction and analysis plans [7–9]. An MA is an SR that uses quantitative methods or statistical techniques to pool the data found in the literature, increasing the effective sample size [7, 8]. For both SRs and MAs, a comprehensive search is necessary to identify all relevant results, including published, unpublished, and gray literature. The more complete the search, the greater the potential for a high-quality study, as shown with a Horizon estimate. SRs and MAs are discussed in more detail in Chap. 16.

A lesser quality RCT, defined as an RCT with less than 80% follow-up, no blinding, or improper randomization, is considered Level II evidence when assessing therapeutic research [10]. Although high-quality RCTs represent the highest LOE, they are not suitable for addressing every surgical research question. When RCTs are unethical, impractical, or adequate resources are not available, researchers may elect to conduct an observational study (e.g., prospective cohort

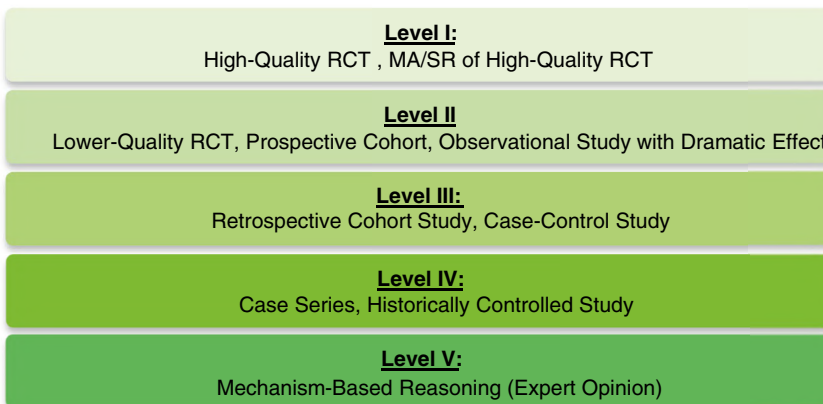
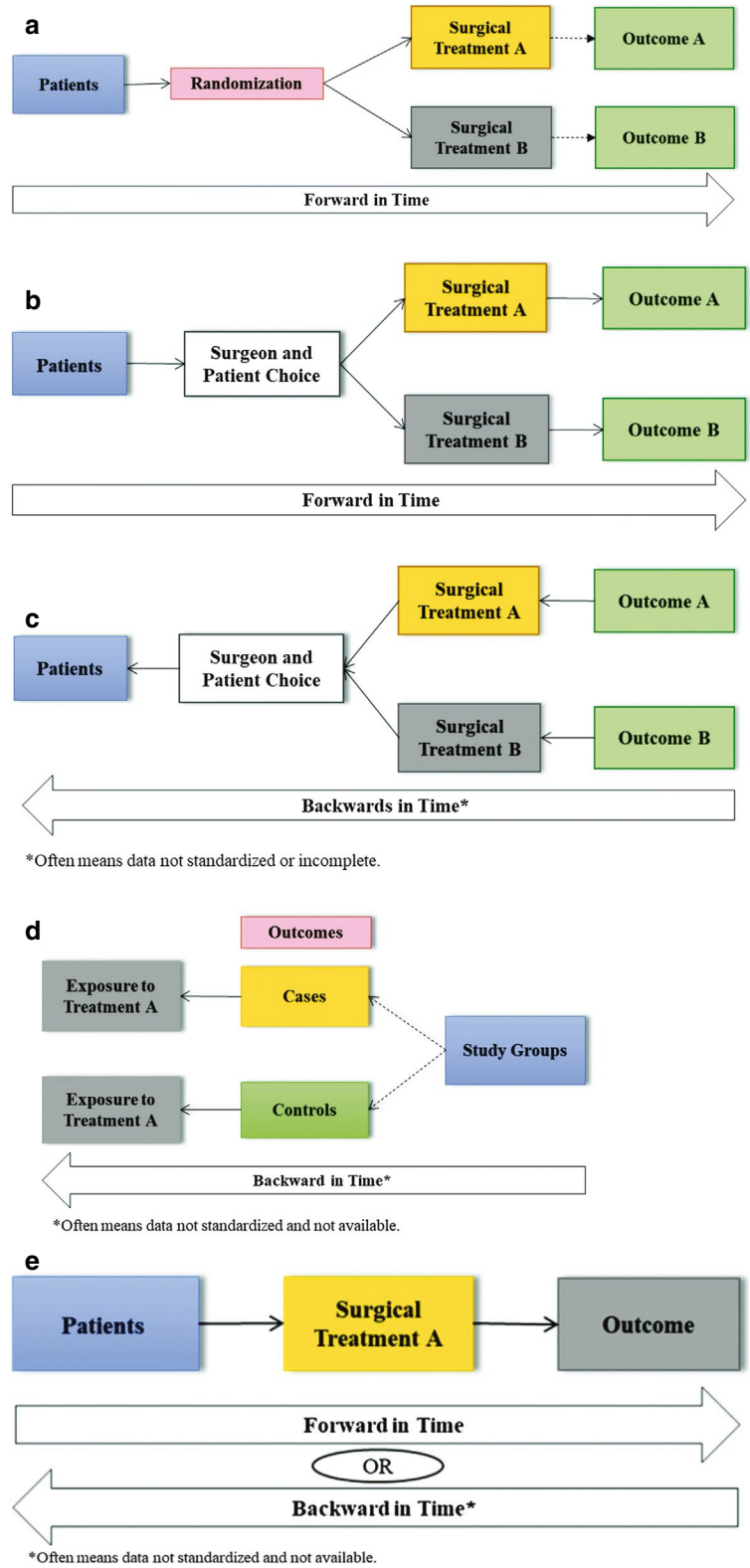


Fig. 5.1 Levels of evidence and their associated study designs for therapeutic research. *RCT* randomized controlled trial. (Created using information from *JBJS* [4] and Elsevier [6])

Fig. 5.2 (a) RCT study design for therapeutic studies. (b) Prospective cohort design for therapeutic studies. (c) Retrospective cohort design for therapeutic studies. (d) Case-control design for therapeutic studies. (e) Case series design for therapeutic studies



study and case-control study) [4]. A prospective cohort study for therapeutic research identifies a group of patients who have received the treatment of interest and compares them with a group of patients who received an alternative treatment (Fig. 5.2b). Patients in both groups are followed for a specified period of time to determine their outcomes [11]. These studies are similar to the RCT in that the criteria for inclusion, the outcomes of interest, and time frame for follow-up are determined prior to the occurrence of any events/outcomes. In contrast to RCTs, the treating surgeon, patient, or both parties, determine which treatment the patient receives, thus introducing a potential source of bias. When a large effect is observed and there are no other considerations for downgrading the quality of the study, the evidence of an observational study is considered satisfactory for Level II. MAs and SRs may also be categorized under Level II evidence if the studies included in the analysis or review are considered Level II evidence, or if the included studies represent Level I evidence but demonstrate inconsistent results [4].

Retrospective cohort studies comparing two or more treatments are commonly encountered in the surgical literature and are classified as Level III evidence. In a retrospective cohort study, the research question is formulated after both the intervention and outcomes have occurred. Researchers review previously recorded data (e.g., from medical records or a registry) to identify patients who received the treatments of interest and then compare their outcomes (Fig. 5.2c) [4]. A case-control study, also at Level III on the hierarchy, compares patients who have the disease or outcome of interest (cases) with controls (patients who do not have the disease or outcome of interest) and looks back in time to compare how frequently the exposure to a risk (e.g., fracture and vitamin D deficiency) is present in each group to determine the relationship between the risk factor and the disease (Fig. 5.2d). The status may not have been recorded in the past, and therefore, the presence or absence of the risk factor cannot be determined. Case-control studies are described in greater detail in Chap. 18.

Case series are frequently reported in the surgical literature and are classified as Level IV evidence [4]. A case series is an expansion of an individual case report to include a group of patients with the outcome of interest (Fig. 5.2e). A case series does not involve a comparison group; however, may be valuable at making observations about a rare surgical case or describing a new surgical technique [4]. A case series is also beneficial in developing hypotheses that are to be tested with analytic studies; however, it is not possible to make causal inferences about the relationship between risk factors and the outcome of interest [8, 12]. Similar to case-control studies, in a case series, the status may not have been recorded in the past, so the presence or absence of risk factors cannot be determined. Lastly, case series may not be generalizable, as they often focus on a single surgeon or center's experiences [8]. Case series are very common in the surgical literature as they are relatively easy and inexpensive to do (see Chap. 19: Case Series).

A historically controlled study is another study design that constitutes Level IV evidence. This study design includes a control and a treatment group which share the same disorder. The treatment group is comprised of individuals who are currently receiving the treatment of interest and the control group consists of individuals who received an alternative treatment in the past. This design allows researchers to investigate the effects of a new treatment quickly and inexpensively. However, this study design is vulnerable to bias and requires that the selection of historical controls be sufficiently similar to that of the treatment group [13]. Lastly, since the status may not have been recorded in the past, the presence or absence of the risk factor cannot be determined.

Lowest on the hierarchy of evidence for therapeutic research is mechanism-based reasoning [4]. Mechanism-based reasoning is based on expert opinion and is intended to generate hypotheses or interpret evidence from other sources. This LOE relies on the knowledge of underlying mechanisms to predict what the relevant effect of a therapy (or diagnostic test or prognosis, depending on the area of research) will be for the patient [14]. This LOE has the

greatest potential for error and is, therefore, the lowest on the hierarchy.

Level I through Level IV evidence can be graded downward based on study quality, imprecision, indirectness, or inconsistency between studies or when the effect size is very small. Similarly, these studies can be graded upward if there is a dramatic effect size [11].

Clinical Example

An example of a Level I evidence study is, “Biologic vs Synthetic Mesh for Single-stage Repair of Contaminated Ventral Hernias: A Randomized Clinical Trial”, which is appraised in Chap. 11.

Levels of Evidence for Diagnostic Research

Diagnostic research evaluates the ability of a diagnostic test to determine the presence or absence of the condition in question. Research questions that diagnostic research addresses include “Is this (early detection) test worthwhile?” and “Is this diagnostic or monitoring test accurate?” The appraisal of a diagnostic test is explained in Chap. 21: The hierarchy of evidence for diagnostic research is listed in Fig. 5.3.

Similar to therapeutic research, RCTs are considered to be Level I evidence. Diagnostic RCTs are randomized comparisons of two diagnostic interventions (one standard and one experimental) with identical therapeutic interventions based on the results of the competing diagnostic interventions (e.g., disease: yes or no) and with the study outcomes being clinically important consequences of diagnostic accuracy (Fig. 5.4a) [15]. Additionally, testing of previously developed diagnostic criteria, using a cohort study design with consecutive patients, consistently applied reference standards, and blinding is also considered to be Level I evidence [4]. In these prospective diagnostic accuracy cohort studies, patients with suspected disease or condition undergo both the diagnostic intervention or criteria being evaluated and the diagnostic reference standard (Fig. 5.4b) [15]. Diagnostic accuracy or the performance of the previously developed experimental diagnostic intervention or criteria is measured using a 2 × 2 table [15]. Diagnostic tests with three or more levels allow for likelihood ratios to be computed, but not sensitivity and specificity. Therefore, diagnostic tests should include likelihood ratios and not sensitivity and specificity, as they go in the wrong direction (see Chap. 21). While diagnostic cohort studies inform us about the relative accuracy of an experimental diagnostic intervention compared to a reference standard, they do not inform us about whether the differ-

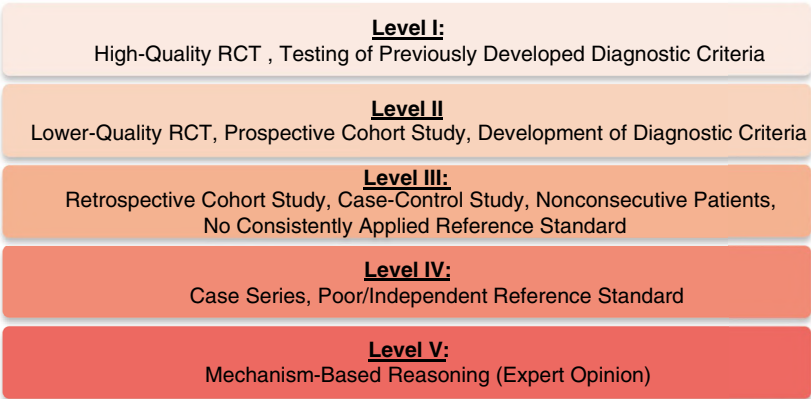


Fig. 5.3 Levels of evidence and their associated study designs for diagnostic research. RCT randomized controlled trial. (Created using information from JBJS [4] and Elsevier [6])

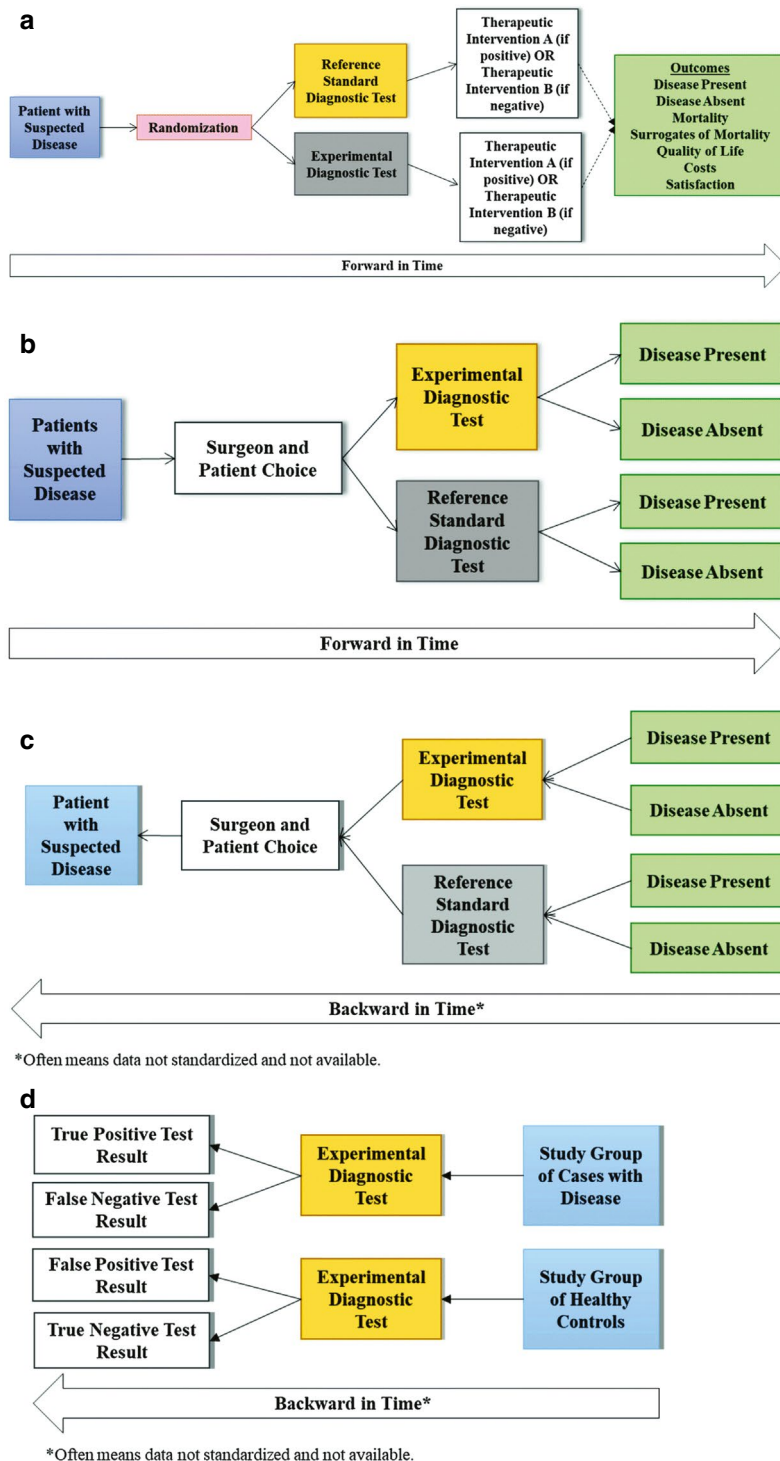


Fig. 5.4 (a) RCT study design for diagnostic studies. (b) Case-control design for diagnostic studies. (c) Prospective cohort design for diagnostic studies. (d) Retrospective cohort design for diagnostic studies. (e) Case series design for diagnostic studies

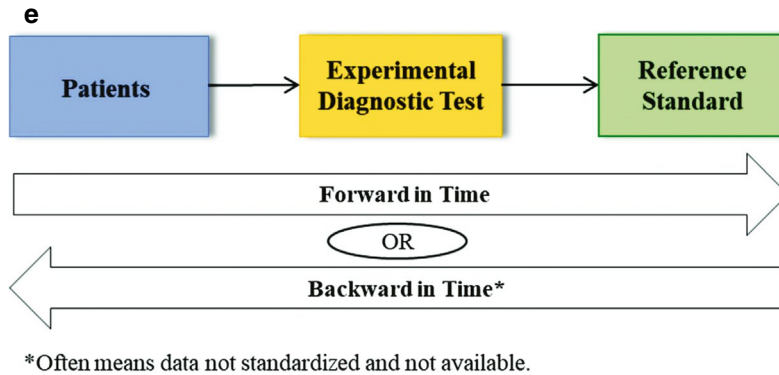


Fig. 5.4 (continued)

ences in accuracy are clinically important, or the degree of clinical importance (in other words, the impact on patient outcomes) [15].

Level II evidence includes a prospective cohort study, as described above, evaluating a new diagnostic test or diagnostic criteria. Additionally, studies that describe the development of a new diagnostic test or criteria, that include consecutive patients with consistently applied reference standards and blinding, are also considered Level II evidence.

Level III study designs, include retrospective cohort studies. These studies are similar to the cohort studies described above, however, data are collected retrospectively (Fig. 5.4c). Diagnostic case-control studies are also considered to be Level III evidence (Fig. 5.4d). Additionally, prospective cohort studies that includes nonconsecutive patients and does not include a consistently applied reference standard are also considered to be Level III evidence, and, therefore, may not be present in records.

As with therapeutic research, case series describing a diagnostic test or diagnostic criteria are considered to be Level IV evidence (Fig. 5.4e). When case series, case-control, or retrospective studies look back in time, the included data may not be standardized, or the data are unavailable and were not recorded in the past, resulting in a lower quality of evidence. Additionally, prospective cohort studies that have a poor or nonindependent reference standard are also considered to be Level IV evidence. Lastly, diagnostic studies using mechanism-based reasoning are considered to be Level V evidence.

Clinical Example

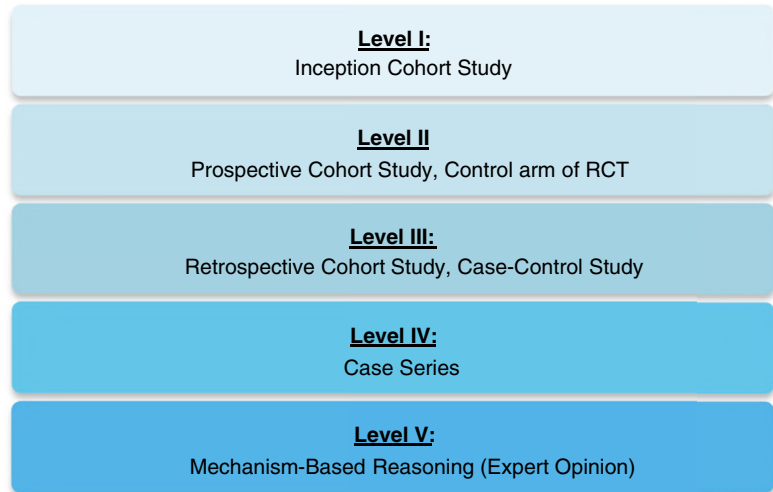
An example of a Level II evidence study is the “Comparison of diagnostic performance of the ACR and Kwak TIRADS applying the ACR TIRADS’ size thresholds for FNA” by Huh et al., which is appraised in Chap. 21.

Levels of Evidence for Prognostic Research

Prognostic research estimates the risk of future outcomes in individuals based on clinical and nonclinical characteristics [10]. It addresses questions regarding the natural history of the disease. By identifying prognostic studies involving patients with a similar clinical presentation, an attending surgeon is able to inform their patients about their expected clinical course and make better-informed decisions on treatment [10]. Typically, prognostic research consists of prospective, observational studies that aim to assess the causes of disease progression, prediction of risk in individuals, and individual response to treatment [16]. As it is not possible to randomly assign these factors, one usually cannot conduct an RCT to address a prognostic research question. Therefore, compared to therapeutic and diagnostic research, prognostic research studies follow different criteria in determining LOE (Fig. 5.5) [17].

Inception cohort studies represent the highest LOE in prognostic research. In an inception cohort study, all patients are enrolled at a com-

Fig. 5.5 Levels of evidence and their associated study designs for prognostic research



mon time early in the development of their disease or condition (near the onset of symptoms, soon after diagnosis, or at detection of a clinically important pathological event), and are followed thereafter [18]. Outcomes and factors associated with patient outcomes are then reported (Fig. 5.6a).

Level II evidence includes a prospective cohort study (as described above) in which patients are enrolled at different points of their disease (Fig. 5.6b). Additionally, the use of data from a control arm of a RCT to determine prognostic outcomes is also considered Level II evidence. Retrospective cohort studies (Fig. 5.6c) and case-control studies (Fig. 5.6d) reporting on prognostic measures are considered Level III evidence.

Similar to therapeutic and diagnostic research, case series are considered to be level IV evidence in prognostic research (Fig. 5.6e) and mechanism-based reasoning is considered to be Level V evidence in prognostic research [7].

Clinical Example

An example of a prognostic study deemed Level III evidence is the “Bicuspid aortic valve repair: outcomes after 17 years” by Jasinski et al., appraised in Chap. 22.

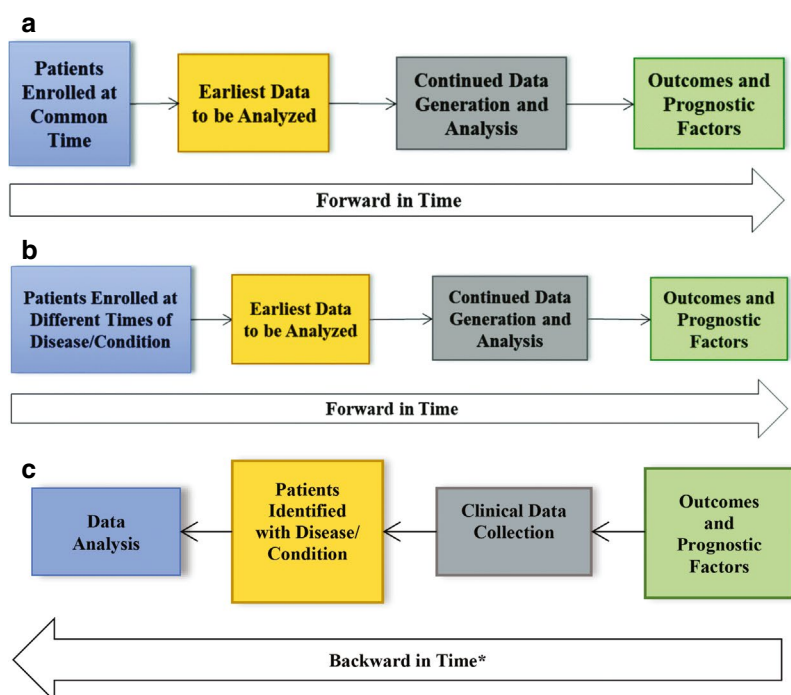
Levels of Evidence for Economic Research

Economic research evaluates whether or not the intervention or treatment offers the greatest benefit for the amount of dollars spent [4]. All LOE for economic research include computer simulation models (CSM), using Monte Carlo simulation or Markov models. The inputs for each LOE differs. For Level I therapeutic evidence, inputs should be derived from Level I studies, as well as include lifetime time duration, outcomes expressed in dollars per quality-adjusted life years (QALYs), and uncertainty examined using probabilistic sensitivity analyses [4].

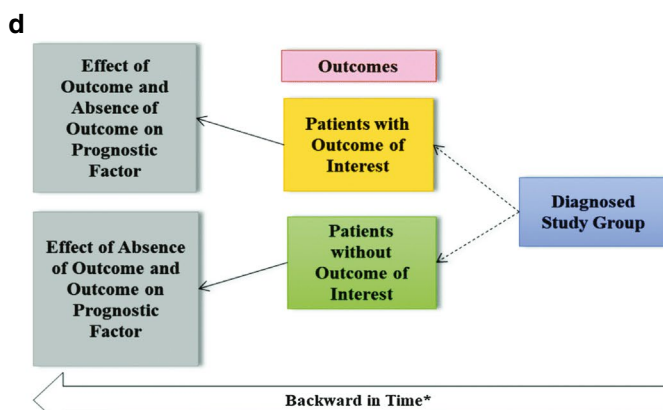
Level II evidence consists of computer simulation from Level II studies, including lifetime time duration, outcomes expressed in QALYs and uncertainty examined using probabilistic sensitivity analyses. Level III evidence also uses inputs derived from Level II studies, however, differs as it uses a relevant time horizon, less than a lifetime, and outcomes expressed in dollars per QALYs and stochastic multilevel sensitivity analyses [4].

Level IV and V on the hierarchy of evidence for economic research uses a decision tree over the short-time horizon. Level IV uses input data from original Level II and III studies and uncertainty is examined by univariate sensitivity analyses. Level V differs as it uses input data informed by prior economic evaluation and uncertainty is examined

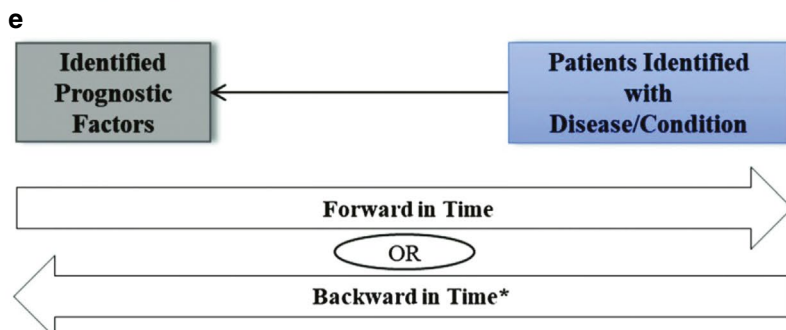
Fig. 5.6 (a) Inception cohort design for prognostic studies. (b) Prospective cohort study design for prognostic studies. (c) Retrospective cohort design for prognostic studies. (d) Case-control design for prognostic studies. (e) Case series design for prognostic studies



*Often means data not standardized and not available.



*Often means data not standardized and not available.



*Often means data not standardized and not available.

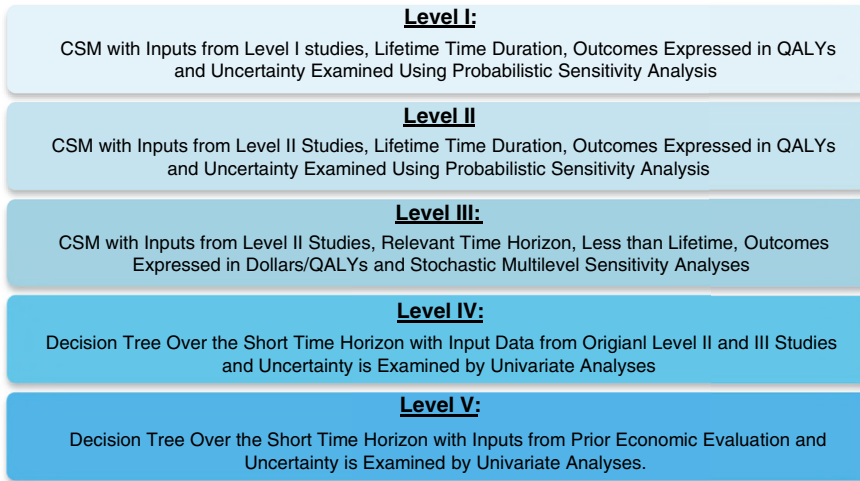


Fig. 5.7 Levels of evidence and their associated study designs for economic research. *CSM* computer simulation model. (Created with information from *JBJS*)

by univariate sensitivity analyses [4]. Figure 5.7 summarizes the LOE for economic research and Chap. 24 provides guidance on how to critically appraise and understand economic evaluations.

prognostic studies, while a computer simulation model with inputs derived from Level I therapeutic studies represent the best available evidence for economic analyses.

Clinical Example

An example of a Level I evidence study is the clinical scenario in Chap. 24, comparing lumbar spinal stenosis (LSS) and X-stop interspinous distractor devices (IDD).

Summary

To effectively practice evidence-based surgery, a clear understanding of the hierarchy of evidence is necessary. In this chapter, we discussed the hierarchy of evidence as it applies to therapeutic, diagnostic, prognostic, and economic research. Of note, SRs and MAs of high-level clinical studies are suitable for evaluating all clinical questions. Questions about therapy and diagnostics are best addressed with high quality, definitive randomized controlled trials. Inception cohort studies are the highest available evidence for

Concluding Remarks

The LOE are a vital component of evidence-based surgical practice. The hierarchy of evidence provides a valuable tool by which surgeons are able to understand and rank the surgical literature by study design. One, however, should not be dismissive of the lower LOE. Early in their infancy novel interventions progress from a case report, to case series, to comparative cohort studies. Once such innovations are adopted widely, they are then compared in RCTs or meta-analyses. The problem arises when in the face of higher LOE surgeons choose to base their practice on lower LOE because they lack the skills to distinguish the two. When faced with multiple sources and variable LOE, the surgeon should be able to separate the “chaff from the wheat.” Later chapters in this book will discuss the next steps of critical appraisal which is specific to each study design.

Appendix: Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence [3]

Question	Step 1 (Level 1 ^a)	Step 2 (Level 2 ^a)	Step 3 (Level 3 ^a)	Step 4 (Level 4 ^a)	Step 5 (Level 5)
<i>How common is the problem?</i>	Local and current random sample surveys (or censuses)	SR of surveys that allow matching to local circumstances ^b	Local non-random sample ^b	Case series ^b	—
<i>Is this diagnostic or monitoring test accurate?</i> (diagnosis)	SR of cross-sectional studies with consistently applied reference standard and blinding	Individual cross-sectional studies with consistently applied reference standard and blinding	Nonconsecutive studies, or studies without consistently applied reference standards ^b	Case-control studies, or “poor or nonindependent reference standard ^b ”	Mechanism-based reasoning
<i>What will happen if we do not add a therapy?</i> (Prognosis)	SR of inception cohort studies	Inception cohort studies	Cohort study or control arm of randomized trial ^a	Case-series or case-control studies, or poor-quality prognostic cohort study ^b	—
<i>Does this intervention help?</i> (treatment benefits)	SR of randomized trials or <i>n</i> -of-1 trials	Randomized trial or observational study with dramatic effect	Non-randomized controlled cohort/follow-up study ^b	Case-series, case-control studies, or historically controlled studies ^b	Mechanism-based reasoning
<i>What are the COMMON harms?</i> (treatment harms)	SR of randomized trials, systematic review of nested case-control studies, <i>n</i> -of-1 trial with the patient you are raising the question about, or observational study with dramatic effect	Individual randomized trial or (exceptionally) observational study with dramatic effect	Non-randomized controlled cohort/follow-up study (postmarketing surveillance) provided there are sufficient numbers to rule out a common harm (for long-term harms the duration of follow-up must be sufficient) ^b	Case-series, case-control, or historically controlled studies ^b	
<i>What are the RARE harms?</i> (treatment harms)	SR of randomized trials or <i>n</i> -of-1 trial	Randomized trial or (exceptionally) observational study with dramatic effect			
<i>Is this (early detection) test worthwhile?</i> (screening)	SR of randomized trials	Randomized trial	Non-randomized controlled cohort/follow-up study ^b	Case-series, case-control, or historically controlled studies ^b	

SR Systematic review

^a Level may be graded down on the basis of study quality, imprecision, indirectness (study of PICOT does not match question PICOT), because of inconsistency between studies, or because the absolute effect size is very small. Level

may be graded up if there is a large or very large effect size

^b A systematic review is generally better than an individual study

References

1. Canadian Task Force on the Periodic Health Examination. The periodic health examination. *Can Med Assoc J.* 1979;121(9):1193–254. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1704686/pdf/canmedaj01457-0037.pdf>. Accessed 06 Mar 2024.
2. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest.* 1989;95(Suppl 2):2S–4S.
3. OCEBM Levels of Evidence Working Group. The Oxford 2011 levels of evidence. <http://www.cebm.net/index.aspx?o=5653>. Accessed 06 Mar 2024.
4. JBJS. Journals Level of Evidence. *J Bone Joint Surg.* <https://journals.lww.com/jbjsjournal/Pages/Journals-Level-of-Evidence.aspx> (Published 2015).
5. Wright JG, Swiontkowski MF, Heckman JD. Introducing levels of evidence to the journal. *J Bone Jt Surg.* 2003;85(1):1–3.
6. Elsevier. Levels of evidence in research. <https://scientific-publishing.webshop.elsevier.com/research-process/levels-of-evidence-in-research/> (2005). Accessed 06 Mar 2024.
7. Sprague S, McKay P, Thoma A. Study design and hierarchy of evidence for surgical decision making. *Clin Plast Surg.* 2008;35:195–205. <https://doi.org/10.1016/j.cps.2007.10.007>.
8. Petrisor B, Bhandari M. The hierarchy of evidence: levels and grades of recommendation. *Indian J Orthop.* 2007;41(1):11–5.
9. Fineout-Overholt E, Melnyk BM, Stillwell SB, Williamson KM. Evidence based practice step by step: critical appraisal of the evidence: part I. Source. *Am J Nurs.* 2010;110(7):47–52. <http://www.jstor.org/stable/pdf/25684627.pdf?refreqid=excelsior%3Aadf45c2c61b6114a9a7283ab8269142e>. Accessed 06 Mar 2024.
10. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ.* 2009;338:b375. <https://doi.org/10.1136/BMJ.B375>.
11. Wupperman R, Davis R, Obremskey WT. Level of evidence in spine compared to other orthopedic journals. *Spine.* 2007;32(3):388–93.
12. Brighton B, Bhandari M, Tornetta P, Felson DT. Hierarchy of evidence: from case reports to randomized controlled trials. *Clin Orthop Relat Res.* 2003;413(413):19–24. <https://doi.org/10.1097/01.blo.0000079323.41006.12>.
13. Viele K, Berry S, Neuenschwander B, Amzal B, Chen F, Enas N, Hobbs B, Ibrahim JG, Kinnersley N, Lindborg S, Micallef S, Roychoudhury S, Thompson L. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat.* 2014;13(1):41–54. <https://doi.org/10.1002/pst.1589>.
14. Howick J, Chalmers I, Glasziou P, Greenhalgh T, Heneghan C, Liberati A, Moschetti I, Phillips B, Thornton H. Explanation of the 2011 Oxford Centre for evidence-based medicine (OCEBM) levels of evidence (background document). Oxford Centre for Evidence-Based Medicine; 2011.
15. Rodger M, Ramsay T, Fergusson D. Diagnostic randomized controlled trials: the final frontier. *Trials.* 2012;13(1):137. <https://doi.org/10.1186/1745-6215-13-137>.
16. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. *BMJ.* 2009;339:b4184. <https://doi.org/10.1136/BMJ.B4184>.
17. Devries JG, Berlet GC. Understanding levels of evidence for scientific communication. *Foot Ankle Spec.* 2010;3(4):205–9.
18. Porta M, Greenland S, Hernán M, dos Santos SI, Last J. For the international epidemiological association. A dictionary of epidemiology. 6th ed. New York: Oxford University Press; 2014.

Evaluating Surgical Interventions

6

Tyler McKechnie, Yung Lee,
Aristithes G. Doumouras, and Dennis Hong

Abbreviations

ANOVA	Analysis of Variance
ARR	Absolute Risk Reduction
CONSORT	Consolidated Standards of Reporting Trials
MeSH	Medical Subject Headings
NNT	Number-Needed-to-Treat
OR	Odds Ratio
RR	Risk Ratio
RRR	Relative Risk Reduction

Introduction

Clinical decision-making in surgery has evolved dramatically in the past several years. Traditionally, surgeons coalesced anecdotal evidence, previous experience, expert opinion,

and extrapolation of basic science into clinical decision-making. While this often served patients well, it is also possible that this decision-making process led to suboptimal treatment. To combat this, a plethora of surgical research has been ongoing for the past several decades to generate, refine, and utilize evidence-based surgical practices [1]. The evidence itself can vary greatly in quality and therefore several grading systems have been developed to assess quality [2, 3]. Regardless of the grading system, the general hierarchy of evidence remains the same and is based on the propensity of a certain study type to introduce bias into its conclusions. Bias, specifically, refers to any systematic deviation from the truth, which could result in an underestimation or overestimation of the true effect of an intervention, and thus threaten the internal validity of a study [4].

The lowest level of evidence after expert opinion consists of retrospective case series which present outcomes without comparison, and retrospective cohorts, which often are used to compare two or more therapies (see Chap. 5). While uncomplicated to design, the retrospective nature of these studies does not allow for the collection of all potential confounding factors. This limits the ability to infer causation from the conclusions. The next level of evidence consists of prospective, nonrandomized controlled trials. While the prospective nature of these studies allows for better data collection, the lack of randomization can still lead to an

T. McKechnie
Division of General Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: tyler.mckechnie@medportal.ca

Y. Lee
Department of Surgery, McMaster University,
Hamilton, ON, Canada
e-mail: yung.lee@medportal.ca

A. G. Doumouras · D. Hong (✉)
Department of Surgery, McMaster University,
Hamilton, ON, Canada

Department of Surgery, St. Joseph's Healthcare,
Hamilton, ON, Canada
e-mail: aristithes.doumouras@medportal.ca

imbalance within groups for unmeasured variables that would not be accounted for fully with statistical analyses (i.e., residual confounding). Properly designed and conducted randomized trials provide a high level of internal validity to their conclusions but can be limited by their generalizability [3]. Accordingly, the highest level of evidence is when several randomized trials looking at the same topic can be combined into a meta-analysis [5, 6]. Meta-analyses combine the results of several trials to boost statistical power and minimize false negative results, while also potentially increasing generalizability. However, one should be cautious of meta-analyses of observational data as *a meta-analysis of biased data is still biased* though these types of studies may be useful in summarizing the current observational evidence for a topic. Ideally, surgeons can evaluate all of the available evidence they have for the clinical problem at hand and provide optimal care based on this assessment, considering effect estimates, risk of bias, and certainty of evidence.

Clinical Scenario

You are a minimally invasive surgery fellow at a large academic institution in the United States. A 50-year-old female presents to the outpatient surgical clinic with clinical and radiographic evidence of a ventral hernia. She has previously undergone an open right-hemicolectomy for cecal volvulus via a midline laparotomy, the cranial portion of which is the source of the ventral hernia. She is quite symptomatic, describing intermittent abdominal pain associated with the hernia and a gradual increase in size, which has prompted her to seek operative repair. However, she has a body mass index of 36 kg/m². As such, you recommend preoperative optimization by way of weight loss and plan to see her again in clinic in 3-to-4 months for reconsideration of elective hernia repair. The patient is satisfied with this plan, but prior to leaving the clinic she inquires more about the prospective surgery. The patient knows that your staff surgeon performs both robotic and laparoscopic ventral hernia

repair. She asks you which of the two approaches you would recommend for her case, or if there is a difference at all. You inform her that both are safe options and that prior to her next visit, you will review the literature to inform an in-depth discussion with her at the time of her follow-up. To better inform your literature review, you ask her which outcomes are most important to her. She states that she wishes to minimize her risk of recurrence, risk of infection, the amount of time spent in hospital, and pain after surgery.

Search Strategy

You design a search strategy that aims to cast a wide net and include all relevant articles while ultimately trying to focus on a much smaller number of key articles. Accordingly, you search MEDLINE with a broad strategy that includes both the Medical Subject Headings (MeSH) “Herniorrhaphy,” “hernia,” ventral,” “laparoscopy,” and “robotic surgical procedures,” and the non-MeSH keywords “hernia repair,” “ventral hernia,” “laparoscopic surgery,” and “robotic surgery.” Using non-MeSH terms ensures the inclusion of all relevant abstracts. These search terms are combined with the “OR” Boolean function to create the initial database. From this, abstracts can be reviewed individually or, if still too numerous, the Boolean “AND” function can be used to narrow down the choices and create a more relevant set. For practical purposes, we can also limit it to the English language though this may introduce bias into systematic reviews. To further narrow down the list, we can also look directly for clinical trials (see Chap. 4). In doing this, we get 13 hits, and of these hits, one is a randomized controlled trial that was recently published by Dhanani and colleagues in the *Annals of Surgery* and reports all of the aforementioned outcomes that are important to your patient [7]. The purpose of this chapter is to assist the reader to interpret the results/data presented in a surgical article. Therefore, we will appraise the Dhanani et al. article and aim to determine if the authors’ conclusions are supported by the data they provide [7].

PICOT Question

The evaluation of every research article begins by identifying the research question to ensure that the findings of the study will be applicable to your given clinical scenario. The PICOT format is the most used method for framing research questions [8]. For the trial by Dhanani et al. the research question as per the PICOT format is as follows [7]:

- Population:* Adult patients with a diagnosis of a ventral hernia less than 12 cm wide on physical examination, who were able to tolerate pneumoperitoneum during laparoscopy, had no previous history of operative treatment for bowel obstruction, and had no active infection.
- Intervention:* Robotic ventral hernia repair with intraperitoneal onlay mesh.
- Comparison:* Laparoscopic ventral hernia repair with intraperitoneal onlay mesh.
- Outcome(s):* Length of stay in the hospital (primary outcome), surgical site infection, surgical site occurrence, hernia occurrence, readmission, reoperation, and mortality.

Time: Length of stay was measured up to 90 days in the hospital. Other outcomes were followed up for 2 years.

Of note, the primary outcome, length of stay in hospital, was reported in a previous publication by the same research group in 2020 [9]. Otherwise, the remainder of the above outcomes were reported in the 2-year follow-up study by Dhanani et al.

Appraisal of This Article

The evaluation of surgical interventions follows the framework shown in Table 6.1. When considering the evidence, there are three main aspects of a study that should be assessed: internal validity, the results, and external validity (Table 6.1). We will discuss each of these points individually and how to assess them. We will scrutinize the results of the Dhanani et al., which, are presented in Table 6.2 [7].

Table 6.1 How to appraise an article evaluating surgical interventions. Adapted from Urschel et al. [10]

Question	Appraisal
Are the results valid?	Was participant assignment randomized, and the randomization process “concealed”? Were trial personnel “blinded” to treatment and apart from the experimental intervention, were the groups treated equally? Were all participants who entered the trial accounted for and was follow-up adequate? Were participants analysed according to the “intention to treat” principle? Was the trial large enough to detect a difference?
What are the results?	What are the outcomes and how were they measured? How large was the treatment effect? How precise was the estimate of the treatment effect?
Are the results applicable to my patients?	Were the trial participants similar to my patients? Were the measured outcomes clinically relevant? Are my surgical skills similar to those of the trial surgeons?

Table 6.2 Outcomes and mortality among trial participants [7]

	Outcome	Laparoscopic hernia repair (n = 47)	Robotic hernia repair (n = 54)	Absolute difference		p-value
				Percentage points	95% CI	
Primary outcomes (90 days)	<i>Length of stay in hospital</i>	Mean 0.38 days (SD 1.17)	Mean 0.41 days (SD 0.84)	0.03	(−0.36 to 0.42)	0.82
Secondary outcomes (2 years)	<i>Hernia recurrence</i>	6 (12.8)	2 (3.7)	−9.1	(−19.9 to 1.7)	0.12
	<i>Surgical site infection</i>	0	0	0	N/A	1.00
	<i>Seroma</i>	4 (8.5)	7 (13.0)	4.5	(−7.5 to 16.5)	0.46
	<i>Hematoma</i>	0	0	0	N/A	1.00
	<i>Readmission</i>	3 (6.4)	0	−6.4	(−13.4 to 0.6)	0.10
	<i>Reoperation</i>	5 (10.6)	0	−10.6	(−19.4 to −1.8)	0.02
	<i>Functional status (modified activities assessment scale)</i>	Mean 76.2/100 (SD 30.3)	Mean 84.1/100 (SD 19.0)	7.9	(−1.6 to 17.4)	0.17
	<i>Pain score (visual analog scale)</i>	Mean 2.3/10.0 (SD 4.6)	Mean 1.3/10.0 (SD 2.3)	1.0	(−0.3 to 2.3)	0.22
Death	<i>Satisfaction score (ventral hernia repair satisfaction)</i>	Mean 9.3/10.0 (SD 1.7)	Mean 9.3/10.0 (SD 1.5)	0	(−0.6 to 0.6)	0.31
		1 (2.1)	0	2.1	(−2.0 to 6.2)	0.47

Appraisal of an Article: Are the Results Valid?

Internal validity refers to the soundness of the methodology of the study and relates directly to our ability to infer causation from the results. In observational studies, the correlations presented in the results may not actually represent a true causal relationship and therefore using their conclusions without proper scrutiny may lead to unnecessary, ineffective, or potentially harmful patient care. Two types of errors can threaten internal validity [11]. Type I errors, or false positives, occur when calling a treatment useful when it is not. Type II errors, or false negatives, occur when concluding a treatment has no effect when it is useful (see Chap. 29). There are numerous examples of observational study con-

clusions that turned out to be equivocal or even harmful when examined in a randomized setting. For example, the use of perioperative beta-blockers for noncardiac surgery was initially found to reduce the risk of major cardiovascular morbidity in observational data, but the subsequent POISE trial (n = 8351) demonstrated an increased risk of perioperative death and stroke in patients receiving metoprolol [12]. Randomized trials provide the most sound experimental design and results that are most likely to be causal. This section will highlight some of the important points to look for when assessing the internal validity of a study. Importantly, some of this information may be in the study protocol, trial registration, or statistical analysis plan rather than the manuscript itself, rather than the paper itself.

Was Participant Assignment Randomized, and Was the Randomization Process “Concealed”?

Relationships between outcomes predictors of interest are subject to confounding. Confounding is the potential for a third variable to influence the relationship between the outcome and predictor, through independent effects on both the predictor and the outcome, thus limiting our ability to infer direct causation (see Chap. 30). For observational studies, many of these confounders can be identified and adjusted with multivariable statistical analyses. One of the advantages of prospective over retrospective studies is the ability to ensure data collection on important confounding factors. However, there are many unmeasurable or unknown confounding factors that cannot be collected and adjusted for (i.e., residual confounders). It is for these factors that randomization is crucial as it is the best way to balance these unknown or unmeasurable factors between groups. The method of randomization (usually a computer program) should be mentioned in the methods of any trial. Another vital aspect of randomization is allocation concealment. This refers to the concealment of the randomization process from the investigators so that it is unknown what group a participant will be randomized into after they are found to be eligible. If not done properly, investigators can intentionally or unintentionally direct participants toward the group they feel is most suitable thus introducing selection bias into the process and losing much of the benefit of randomization. In the trial by Dhanani et al. randomization was performed with a 1:1 allocation ratio using a computer-generated randomization module with variable blocked randomization [7]. On the day of surgery, surgeons called research assistants who opened sequentially numbered opaque sealed envelopes. The specific timing on the day of surgery is not explicitly stated, however it can be assumed that there was somewhat of a significant interval between allocation and the start of the operation because of the different equipment required for these two approaches. As such, randomization was likely concealed from surgeons until a few hours before the operation. Given that the timing of an operation for a given participant was unlikely to be altered on the day of

surgery, the allocation concealment in this trial was likely adequate and protected against selection bias. The randomization according to a variable blocked schema rather than randomizing single individuals is important as it allows for more balance of possible confounding factors. The variable block sizes also protect against selection bias by limiting investigator ability to predict the randomization schedule and place participants such that the intervention is favoured.

Were Trial Personnel “Blinded” to Treatment and Apart from the Experimental Intervention, Were the Groups Treated Equally?

Another major methodological concept is blinding. This is often confused with allocation concealment, but they are distinct concepts. Blinding refers to ensuring that stakeholders in the trial do not know what treatment participants received as knowing this can influence the behavior of participants and investigators. While blinding of both, participants and investigators is relatively straightforward in drug trials, it is often not possible in surgical trials as surgeons will often need to know the treatment the patient had. One way to minimize this issue in surgical trials is to have different clinicians provide the postoperative care to participants. In the trial by Dhanani et al. [7], after randomization and revelation of concealment, both the research coordinator and the operating surgeons were unable to be blinded given the nature of the intervention. However, to minimize the resultant bias, the participants and postoperative outcome assessors were blinded to the participant's treatment arm. Considering the trial question, this was probably the strongest level of blinding possible.

Were All Participants Who Entered the Trial Accounted for and Was Follow-up Adequate?

Another major issue with trials of all types is attrition. Readers should be concerned if not all participants are accounted for at a trial's conclusion. If a large proportion of participants are unac-

counted for at the end of a trial, the benefits of randomization may be lost. Moreover, bias can be introduced if the dropout is related to some aspect of the procedure itself. If the dropout was random, then the benefits of randomization should be maintained. Therefore, a full report of participant attrition is required for trial interpretation. In addition, the follow-up should be rigorous, blinded, and equal between groups to ensure that all adverse effects are accounted for. It should also be assessed similarly between groups and the follow-up period should be long enough to ensure that the outcomes of interest can manifest themselves. In the Dhanani et al. [7] trial, the follow-up was thoroughly conducted by outcome assessors who were blinded to the treatment arm. Researchers should report their loss to follow-up and the associated reasons for loss to follow-up in the form of a Consolidated Standards of Reporting Trials (CONSORT) flow diagram [13]. Overall, 18.5% of participants were lost to follow-up. In the laparoscopic surgery group, 20.3% were lost to follow-up and in the robotic surgery group, 16.9% were lost to follow-up. Thus, loss to follow-up, while a significant proportion of the overall sample, was similar between groups. Unfortunately, there is no previously established threshold for what constitutes “too much” loss to follow-up. Traditionally, less than 5% loss to follow-up was considered a small percentage and greater than 20% of loss to follow-up was considered a large percentage. A more nuanced approach to understanding the importance of participants lost to follow-up takes into account the percentage lost to follow-up, the type of outcome, and the risk of the event. In this trial, in which most clinical outcomes were dichotomous, and their risks were relatively low, the likelihood that the analyses were biased due to missing participants is relatively high. Importantly, median duration of follow-up was similar between groups. Moreover, the duration of follow-up was appropriate for most of the clinical end points, except for recurrence. Therefore, despite a significant amount of loss to follow-up that may threaten prognostic balance between the two groups and subsequently bias the results, as well as limit the statistical power of these analyses, the duration of follow-up is reassuring.

Were Participants Analyzed According to the “Intention-to-Treat” Principle?

The intention-to-treat principle is also fundamental to ensuring causal inference of results. This principle states that participants should be analyzed in the groups they were originally allocated to, regardless of the treatment received. This is vital in surgical trials as participants of poor operative status sometimes may not receive surgery, despite being randomized for surgery. If these participants were to be analyzed in the non-surgical group, they can bias the results by having healthier people in the surgical group. Sometimes, this principle can lead to issues of validity of conclusions if there are too many participants that did not receive the treatment but in most cases the strategy is sound. In addition, in surgical trials, if there are many participants not receiving the treatment it may provide a pragmatic answer as to the feasibility of the treatment. In the Dhanani et al. [7] trial, analyses were completed by three investigators who were not involved in recruitment or data collection according to intention-to-treat principles. However, the authors do not mention how missing data were handled. This is an issue because the number of missing participants at 2-year follow-up is significant, and thus if a complete case analysis including data only from participants completing follow-up, the analyses were likely biased due to the high likelihood of nonrandom missingness of these data. Moreover, the large proportion of missingness calls into question the statistical power of these analyses.

Was the Trial Large Enough to Detect a Difference?

Ensuring an adequate sample size is essential to answering any clinical question. A randomized trial should clearly describe an a priori sample size calculation and what factors they used to determine the sample size. Generally, calculating a sample size requires the rate of type I error (usually 5%), rate of type II error (usually 20%,

also known as 80% power), the allocation ratio between groups (e.g., 1:1, 2:1), the expected effect and the variance of that effect. While the first three are standard in the sample size calculation, the last few variables can be controlled by investigators. A larger a priori expected effect means a smaller sample size but if that effect is unreasonable then it could lead to an underpowered trial. Conversely, the effect size chosen should also be large enough to be clinically relevant. Power is the complement of the type II error rate for a trial, and can be described as the chance for a trial to produce a false negative result. Therefore, before a negative result can be truly established, the sample size and statistical power should be scrutinized and found to be adequate (see Chap. 29). In addition, the conclusions of other outcomes assessed, for which there was no sample size calculation, should not be assumed to be adequately powered unless explicitly stated by the authors that the trial was powered for secondary outcomes. In the Dhanani et al. trial, the sample size was calculated at 60 participants within each group [7]. This would give the trial a power of 80% and a significance level of 5%, in addition to accounting for a 10% drop out rate, to detect a one-day difference in the number of days spent in hospital within 90 days of the index ventral hernia repair. The investigators randomized 124 participants, thus meeting their a priori sample size goal. However, in this 2-year follow-up, there were only 101 participants that remained for analysis and all of the outcomes were secondary outcomes. As such, this trial was not designed to compare laparoscopic and robotic surgery for the included outcomes (i.e., recurrence, surgical site infection, and seroma) with adequate statistical power. The resultant data presented in this trial were subsequently underpowered and do not provide definitive evidence as to the comparable effectiveness of laparoscopic and robotic approaches to ventral hernia repair in terms of these outcomes. These data should be regarded as hypothesis generating.

What Are the Results?

The key findings of the Dhanani et al. trial are found in Table 6.2 [7]. The 9.1% absolute decrease in hernia recurrence at 2-years in the robotic surgery group, while not statistically significant, does likely represent a clinically meaningful difference. Additionally, the 10.6% decrease in reoperation at 2-years in the robotic surgery group is both clinically and statistically significant. As mentioned previously, however, this trial was not adequately powered for these outcomes and thus the resultant interpretation should be that further study is required to explore these potentially meaningful differences, but definitive conclusions as to the comparable effectiveness of these two surgical approaches are not appropriate.

What Outcomes Were Used and How Were They Measured?

Because of the wide variety of outcomes that can be assessed and the use of sophisticated statistical analyses, simply understanding what the results are can be a challenge for a surgeon. This section will provide a brief summary of how the most common outcomes are reported.

Binary Versus Continuous Outcomes

Most outcomes in the surgical literature are measured as either binary or continuous variables. Binary outcomes represent dichotomous occurrences where patients either have an outcome or they do not (e.g., death, recurrence of hernia). Continuous outcomes represent data that are measured by real numbers such as weight, cost, or length of stay. The trial by Dhanani et al. had a continuous primary outcome (i.e., length of stay in hospital in days) [7]. Their clinical secondary outcomes were all binary (e.g., hernia recurrence, and surgical site infection) and their patient-reported secondary outcomes were ratio continuous variables.

Univariate Testing: Univariable and Multivariable Analyses

There is considerable confusion in the literature as to the nomenclature for testing let alone the actual tests themselves. Univariate testing refers to statistical tests with a single response variable per observation. This represents the vast majority of tests in the surgical literature where a single patient will have a single outcome associated with them. Multivariate analyses refer to when a single patient has multiple outcomes associated with them and are rarely used (i.e., a patient has several weight changes over the course of the trial) [14]. Univariate testing can take the form of univariable or multivariable tests. Univariable tests occur when the outcome is tested against only a single predictor. Examples include the chi-square test for dichotomous data and a *t*-test for continuous data. For observational data, these usually represent preliminary tests and are used to give some exposition to the data rather than as actual conclusions. Alternatively, for randomized trials, these may represent the final analysis because randomization decreases the need for multivariable analysis through balanced prognosis. Multivariable analyses use multiple predictors to explain the outcome of interest. The simplest examples are linear regression for continuous data and logistic regression for binary data. These studies demonstrate the effect of a single predictor of interest while accounting for impact of the other factors used in the model. These analyses, therefore, account for the effect of many different potential confounders which is not done in univariable analyses. In the trial by Dhanani et al. [7], binary outcomes were assessed with a Poisson regression adjusting for randomization and stratification variables. This is a multivariable analysis that analyzes count data. If a “0 cell” is included (i.e., no outcome events), then regression analyses are not possible. In this case, this trial utilized a Fisher Exact test. Lastly, an analysis of variance (ANOVA) was used to analyze patient-reported outcomes. An ANOVA is a univariate test for continuous outcome data, and in general is used when comparing more than two groups. In general, it will produce similar findings to a student’s *t*-test when used to compare two groups. The analyses used in this trial are appropriate.

Multivariable Regression Results: Odds Ratios (OR) and Risk Ratios (RR)

Every surgeon has seen the results of a multivariable regression but correctly interpreting the results can be difficult. For linear regressions, the conclusions are represented by the effect on the outcome of interest by a one unit change in the predictor. For example, if an outcome was weight and the predictor was age, the results would be represented as the amount the weight has changed for each year in age. For dichotomous data analyzed by logistic regression, results are represented by odds ratios. Odds ratios are related to risk ratios but are less intuitive. Their predominant use stems from the fact that they are much easier to calculate within statistical models. Risk ratios (also known as relative risks or the incidence rate ratios) represent the ratio between the actual event rate between groups. For example, if the event happened 60% of the time in group A and 40% in group B then the risk ratio would be 1.5. Similarly, odds ratios are the proportion of the odds of occurrence between the two groups. It is not vital to truly understand the difference between odds and risks, but it is important to know that odds ratios approximate risk ratios when outcomes are rare but *overestimate* risk when the outcomes are common (>10%). Relative risks are common in randomized trials where multivariable logistic regression analyses are used. In the trial by Dhanani et al., comparisons of dichotomous outcome variables were reported in terms of relative risk and comparisons of continuous outcome variables were reported as mean difference, both with their associated 95% confidence intervals. Table 6.3 explains the measures that are used to explain the magnitude and precision of the treatment effect of a surgical intervention. The Dhanani et al. article is explained through these measures in Column 3 with respect to the hernia recurrence outcome [7].

How Large Was the Treatment Effect?

Importantly, the difference between absolute and relative measures needs to be understood by the surgeon contextualizing any results. Odds ratios and risk ratios represent a percentage change

Table 6.3 Terms used to show the magnitude and precision of the treatment effect using hernia recurrence from Dhanani et al. as an example

Term	Description	Example from Dhanani et al. [7]
Risk	The probability of an event occurring is calculated as the number of events of interest divided by the total number within the group	Risk in robotic repair: $2/54 = 0.037$ (3.7%) Risk in laparoscopic repair: $6/47 = 0.128$ (12.8%)
Risk ratio (RR)	The ratio of the risk between two different groups	Risk in robotic repair 0.037 versus laparoscopic repair 0.128 RR: $0.037/0.128 = 0.29$
Odds	The probability of an event occurring is calculated as the number of events of interest divided by the number of participants not experiencing the outcome of interest	Odds in the robotic repair: $2/(54-2) = 0.038$ Odds in the laparoscopic repair: $6/(47-6) = 0.146$
Odds ratio	This is a ratio comparing the probability of the event occurring to the probability of the event not occurring within the same group	Odds in robotic repair 0.038 versus in laparoscopic repair 0.146 OR: $0.038/0.146 = 0.26$
Absolute risk reduction	Absolute reduction in events in one group compared with the other	Complication in robotic repair 3.7% versus in laparoscopic repair 12.8%. ARR = $12.8-3.7 = 9.1\%$
Number needed to treat	The number of patients who need to be treated to prevent one additional outcome. It is defined as the inverse of the absolute risk reduction	ARR from article: 9.1% NNT = $1/ARR = 1/0.091$ NNT = 11 patients
Relative risk reduction	This is the difference in the risk of event between control and experimental group and is calculated as the event rate in the treatment group subtracted from the event rate in the control group, divided by the event rate in the control group. It is the complement of relative risk, expressed as a percentage	Recurrence in laparoscopic surgery 12.8% versus in robotic surgery 3.7%. RRR $(12.8-3.7\%)/12.8\% = 71.1\%$
95% confidence interval	With repeat sampling of studies, 95% of the intervals will include the true value, while a single study does or does not contain the true value	Various methods of calculation

from the occurrence of an event. The absolute risk represents the change of the occurrence of an event on an absolute scale. From a clinical perspective, the latter is usually the more important measure. For example, a study may report a predictor increases the event rate by 30% (risk ratio 1.30). This number seems large but if the event only occurs 1% of the time, then the absolute risk only goes up to 1.3%. The same study could have reported a 0.3% increase in absolute risk—a much less provocative number. Conversely, a study could report a 5% relative increase (risk ratio 1.05) but if the event occurs 50% of the time, then the absolute risk increase is 2.5%, which may be relevant depending on the outcome and the patient population. Despite the clinical utility, absolute risks are usually only reported in randomized trials because they are much more

difficult to model in regression analyses. Accordingly, surgeons must keep in mind the absolute risk of an event to contextualize the results of many trials. The trial by Dhanani et al. [7] does report the absolute risk difference, which for the recurrence outcome is 9%. This was not seen as statistically significant. This, and calculations for the other results from the Dhanani et al. trial are shown in Table 6.2 [7]. The number needed to treat (NNT) of 11 (see column 3 Table 6.3) is an important measure which adds context to our interpretation of the data. It tells us that we have to treat 11 patients with ventral hernia via the robotic method rather than laparoscopic method to avoid one hernia recurrence. Most surgeons may not see this as a huge benefit and may opt for laparoscopic surgery, which is a lower cost alternative.

How Precise Was the Estimate of the Treatment Effect?

The last thing a surgeon should examine are the measures of statistical significance and precision. The statistical significance is represented by p values. These represent the probability of obtaining the results if the null hypothesis (i.e., the assumption that the predictor has no effect) were true. Generally, when this chance is less than 5% (<0.05) we consider the result statistically significant (i.e., that the likelihood that this outcome would occur if the null hypothesis were true is so small, that it is not likely to be explained by chance alone). It should be noted that this specific threshold is completely arbitrary and there is no magical difference between a p value of 0.051 and 0.049. This threshold exists to limit the rate of type I error (e.g., false positives) to 5% (see Chap. 27). Lastly, even if the p value is below this threshold, statistical significance does not equate to clinical relevance. In addition to p values, confidence intervals exist to better characterize the precision of the result. The 95% confidence interval is routinely used and can be interpreted as the interval in which the true effect lies 95% of the time if the same trial with the same sample was repeated (see Chap. 28). Therefore, in studies with small sample sizes, the confidence interval can be quite large but with increasing sample size, the confidence interval becomes narrower. In the Dhanani et al. trial, the two-sided 95% confidence intervals were calculated and for the outcome of difference in hernia recurrence, this interval varied from 0.06 to 1.39 for relative risk, which equated to 95% confidence intervals for absolute difference ranging from -19.9% to 1.7% [7]. One way to interpret this interval is by saying that if this same trial was repeated many times, the true rate difference would lie within this interval 95% of the time.

Are the Results Applicable to My Patients?

External validity refers to the applicability of results to other groups of patients [15]. After the results are understood and the internal validity

assessed, this aspect of any trial should be clearly investigated by the reader and the questions from Table 6.1 should be asked. The most straightforward method of assessing external validity is to compare the PICOT question (i.e., inclusion, exclusion criteria) to the patient in front of you and determine the likeness between the two. If the patient you are treating would fit within the PICOT question of the trial without any important differences, then the results are likely applicable.

Were the Trial Participants Similar to My Patients?

Comparing the patient population of a trial to the patient in front of you is essential to the application of evidence to surgical practice. Many trials have restricted criteria for enrolment and the results may not be directly applicable to the patient you are treating. In addition, they may only be enrolled from a specific patient group (e.g., veterans, men, uncomplicated surgical problems) which may not be the patient in front of you. Often, the patients we treat are older, have more comorbidities, and complex surgical problems than a trial would allow. These factors may mitigate the expected benefit of a treatment and thus the surgeon should know how the evidence relates to the patient in front of them before making any treatment decisions. The protocol for this trial clearly outlines the inclusion and exclusion criteria. Specifically, it included patients 18 years of age or older with a ventral hernia less than 12 cm in width, without a history of an open abdomen or an extensive lysis of adhesions for bowel obstruction, and no active intra-abdominal infection. Moreover, patients had to be deemed appropriate for minimally invasive approaches and to be able to tolerate pneumoperitoneum. Exclusion criteria included patients who did not speak English or Spanish, were deemed unlikely to follow-up, and were judged by their surgeon as unlikely to survive beyond 2 years postoperatively. Additionally, all of these patients were presenting to one of two large centres in Texas, United States of America. If we compare the trial

by Dhanani et al. to the patient we saw in the minimally invasive surgery clinic, we can see that the results of this trial may generalize well to this patient [7]. She is a 50-year old female with a body mass index of greater than 30 kg/m² presenting with an incisional hernia. In the Dhanani et al. trial, the mean age was 49 years, 68.5% of participants were female, 69.4% had a body mass index of greater than 30 kg/m², and 80.6% of hernias were incisional hernias [7]. Moreover, the patient in the clinical scenario lives in the United States, and the Dhanani et al. trial was set at two large hospitals in the United States [7].

Were the Measured Outcomes Clinically Relevant?

Another area that surgeons should question before accepting a “superior treatment” is the choice of main outcomes considered in the trial. The choice of main outcomes should be relevant to both the surgeon and the patient and should be clinically meaningful. Certain biochemical markers may be relevant to surgeons but of no relevance to patients while postoperative pain and the return of function may be less important to surgeons but of great substance to patients. The main outcome of this 2-year follow-up of the trial was recurrence rate at 2 years [7]. Of note, this was different than the primary outcome of the trial, which was length of stay in hospital within 90 days of index operation, which was reported in a previous publication. While a 2-year follow-up is relatively long, many hernia recurrences can occur after this (>5 years) and therefore the long-term durability of this treatment cannot be determined based on the results of this trial [16]. Nonetheless, these are patient important outcomes and thus may be applied in the clinical setting with ease.

Are My Surgical Skills Similar to Those of the Trial Surgeons?

Surgeons should also evaluate whether the treatment itself is feasible within their own practice. Trials on robotic surgery or expensive/difficult to

obtain materials may have little relevance to a surgical practice, especially in low-resource settings, however they may be applicable to larger academic practices in first world countries. Moreover, certain surgical techniques may be beyond the technical proficiency of an individual surgeon and thus the treatment effect would largely be lost in the hands of that surgeon. If clearly published evidence does favor a certain treatment that is available but the surgeon does not perform it well, then the surgeon is faced with three options: proceed with another operation that the surgeon performs well while discussing all options for the patient, refer the patient to a colleague who performs the evidence-based treatment, or seek additional training to master the new technique. Careful thought as to which of these options surgeons pursue is especially required in the context of surgical trials involving novel technology, such as robotic surgery.

Surgical proficiency creates a dilemma for surgical trials utilizing complex procedures and provides a major difference between surgical and medical trials. If the trial uses inexperienced surgeons, it could bias the results away from the treatment, even if it does have a benefit in experienced hands. However, if the trial only utilized highly skilled surgeons, the result may not be applicable to the larger surgical community. The trial by Dhanani et al. utilized robotic surgery as the experimental arm, which is a technique that is not necessarily ubiquitous and does require specialized training [7]. All operations that were included in this trial were performed by three high-volume hernia surgeons with subspecialty training in minimally invasive surgery. Prior to initiating the trial, each of the included centres had completed a minimum of 50 “standardized” repairs to ensure homogenous technique throughout the trial. Given the variability and lack of gold-standard in specific surgical techniques for repairing ventral hernias, the “standardized” technique employed in the trial by Dhanani et al. likely lacks generalizability [7]. Moreover, this trial only pertains to centres in which robotic surgery is available and surgeons use this platform in high volumes. The external validity of this trial is limited, and results should only be applied

within the specific contexts that the trial was conducted. Luckily for our patient that we saw in our minimally invasive surgery clinic, this trial generalizes well to her scenario.

Resolution of the Clinical Scenario

A careful critique of this article demonstrates that the internal validity is fairly high, and therefore the results are likely valid for their stated trial objectives. However, it must be kept in mind that all clinical and patient-reported outcomes in this 2-year follow-up study are secondary outcomes. The trial was not adequately powered to provide definitive conclusions as to the comparable effectiveness of laparoscopic and robotic surgery for outcomes such as hernia recurrence, surgical site infection, and postoperative pain. These data are simply hypothesis generating. Moreover, the external validity of this trial is limited given the highly specialized training of the surgeons performing the operations as well as the expensive and advanced technology used (i.e., robotic surgery platform). Nonetheless, the patient that you saw in the minimally invasive surgery clinic fits well within the inclusion criteria and the demographic observed in the present trial sample. So when she returns to clinic 4 months later, having lost 20 lbs and now having a body mass index of 33 kg/m², you present to her the data you found from the trial by Dhanani and colleagues [7]. You tell her that based on your review of this trial, there are no data to suggest that postoperative length of stay is different between these approaches, nor were there any data to suggest a statistically significant difference in the other outcomes that were important to her (i.e., recurrence, risk of infection, and pain after surgery). Yet this was a small trial and we cannot be confident in the findings with regards to the recurrence, infection, and pain outcomes. The risk of recurrence may have a clinically relevant reduction when robotic surgery is used (i.e., 9.1%), but bigger studies would need to be conducted to confirm this. Overall, she is reassured that both options appear to be safe and effective for repairing her hernia.

Appendix 6.1 Complete Search Strategy Performed on March 1st, 2024 (Medline and Embase Database Example)

1. Herniorrhaphy/
2. Hernia repair.mp.
3. Hernia, Ventral/
4. Ventral hernia.mp.
5. Laparoscopy/
6. Laparoscopic surgery.mp.
7. Robotic Surgical Procedures/
8. Robotic surgery.mp.
9. Or/1-2
10. Or/3-4
11. Or/5-6
12. Or/7-8
13. 9 and 10 and 11 and 12

References

1. Garas G, Ibrahim A, Ashrafian H, Ahmed K, Patel V, Okabayashi K, Skapinakis P, Darzi A, Athanasiou T. Evidence-based surgery: barriers, solutions, and the role of evidence synthesis. *World J Surg.* 2012;36:1723–31.
2. Baker A, Young K, Potter J, Madan I. A review of grading systems for evidence-based guidelines produced by medical specialties. *Clin Med.* 2010;10:358–63.
3. Kavanagh BP. The GRADE system for rating clinical guidelines. *PLoS Med.* 2009;6:e1000094.
4. Higgins J, Green S, editors. 8.2: “Bias” and “risk of bias.” In: *Cochrane handbook for systematic reviews of interventions* version 5.10. 2011. www.handbook.cochrane.org. Accessed 28 Mar 2024.
5. Dickersin K, Scherer R, Lefebvre C. Systematic reviews: identifying relevant studies for systematic reviews. *BMJ.* 1994;309:1286–91.
6. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol.* 2011;64:401–6.
7. Dhanani NH, Olavarria OA, Holihan JL, Shah SK, Wilson TD, Loor MM, Ko TC, Kao LS, Liang MK. Robotic versus laparoscopic ventral hernia repair. *Ann Surg.* 2021;273:1076–80.
8. Riva JJ, Malik KMP, Burnie SJ, Endicott AR, Busse JW. What is your research question? An introduction to the PICOT format for clinicians. *J Can Chiropr Assoc.* 2012;56:167–71.
9. Olavarria OA, Bernardi K, Shah SK, Wilson TD, Wei S, Pedroza C, Avritscher EB, Loor MM, Ko TC, Kao

- LS, Liang MK. Robotic versus laparoscopic ventral hernia repair: multicenter, blinded randomized controlled trial. *BMJ*. 2020;370:m2457.
10. Urschel JD, Goldsmith CH, Tandan VR, Miller JD. Users' guide to evidence-based surgery: how to use an article evaluating surgical interventions. Evidence-based surgery working group. *Can J Surg*. 2001;44:95–100.
 11. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-based medicine working group. *JAMA J Am Med Assoc*. 1993;270:2598–601.
 12. POISE Study Group. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. *Lancet*. 2008;371:1839–47.
 13. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c332.
 14. Hidalgo B, Goodman M. Multivariate or multivariable regression? *Am J Public Health*. 2013;103:39–40.
 15. Sedgwick P. External and internal validity in clinical trials. *BMJ*. 2012;344:e1004.
 16. Burcharth J. The epidemiology and risk factors for recurrence after inguinal hernia surgery. *Dan Med J*. 2014;61:B4846.

A Primer on Outcome Measures

7

Joy MacDermid

Abbreviations

OM	Outcome Measure
PROM	Patient-Reported Outcome Measure
PBOM	Performance-Based Outcome Measure
CBOM	Clinician-Based Outcome Measure
SR	Systematic Review
COSMIN	Consensus-based Standards for the selection of health Measurement Instruments
AMSTAR	A Measurement Tool to Assess systematic Reviews
PROSPERO	International Prospective Register of Systematic Reviews
PICO	P = population, I = intervention, C = control, O = outcome, T = time horizon
AI	Artificial Intelligence

Clinical Scenario

A surgical practice shifted consultations to an online format during the COVID-19 pandemic. Now that practice has returned to in-person care there are patients who for reasons of distance, lack of mobility, or caregiving responsibilities request a virtual consultation. Both personal experience and survey studies have indicated that patient experience is good with virtual consultation [1]. However, the team has traditionally recorded patient-reported outcome measures and physical measurements (ROM, strength, and hand function) as part of their consultation. There is a concern that physical performance outcome monitoring (tests of hand function) is lost during virtual consultations, and as a result the quality of care has been compromised for some patients. This was particularly true for patients with neurological conditions where changes in motor functioning might indicate nerve damage or recovery. The current methods for delivering virtual consultation include talking to patients through a phone call or video/web-based system depending on the abilities and access of the patient. The team is wondering what research has been in validating performance-based assessments through virtual care.

Critical appraisal of evidence on outcome measures (OMs) is particularly challenging since the purposes for the search and appraisal can be multiple. Patient-reported outcome measures (PROMs), performance-based outcome measures

J. MacDermid (✉)

Roth | McFarlane Hand and Upper Limb Centre,
St Joseph's Health Centre, London, ON, Canada

School of Physical Therapy and Department
of Surgery, Western University, London, ON, Canada
e-mail: jmacderm@uwo.ca

(PBOMs), and clinician-based outcome measures (CBOMs) provide different perspectives, benefits, and limitations. Clinical questions may include: What OM are available for a specific construct? What are patient preferences with respect to different OM? What are current practice patterns and the use of an OM in research or clinical practice? What are the comparative theoretical or content differences between different OM? What are the comparative psychometric properties of competing OM? What are the psychometric properties of a particular OM of interest (e.g., reliability, validity, responsiveness, structural validity, ceiling effects/targeting)? What are the impacts of OM on the processes of care and clinical outcomes? What has been recommended by professional group or core consensus process to be used as routine practice (for either research or clinical practice)?

The nature of the clinical questions in evidence-based practice determines the search strategy and the type of research studies that answer the question. Where the question involves an overarching understanding of OM content and measurement properties, an evidence synthesis is ideal. There are many different types of evidence synthesis [systematic reviews (SRs), scoping reviews, integrative reviews, qualitative reviews, etc.]. In this chapter, we will focus on PBOMs and given that we want to identify a range of PBOM we know that either a scoping or systematic review will be required.

Our clinical question is *What PBOM of upper extremity functioning are valid for virtual assessment of patients with neurological conditions?*

Search and Critical Appraisal

The search strategy and retrieval are described in Appendix 1. We focused on evidence from the past 5 years due to changes in the technology and health care delivery in that time window. We reviewed the retrieval nine studies, from which eight did not address our specific question (list and rationale for exclusion in Appendix 1). Based on title and abstract, we were able to find a single SR that addressed our topic [2]. This was a 2024

SR published by Hao et al.; “*Performance-based outcome measures of upper extremity in virtual reality and telerehabilitation: A systematic review*”.

The nature of the research studies being evaluated determines which critical appraisal tools are most appropriate. There are a variety of tools for assessing primary OM studies, with the Consensus-based Standards for the selection of health Measurement INstruments (COSMIN) being comprehensive supports for conducting and evaluating primary studies and evidence synthesis of OM [3–6]. Positive aspects of COSMIN include standardization of processes, implementation tools including critical appraisal tools, rigorous attention to evolving theory and evidence in psychometrics, widespread use, and thorough attention to multiple domains of OM. Negative aspects of COSMIN include complexity which may compromise reliability especially in nonexpert reviewers, and that some criteria are based on expert opinion and may be considered dogmatic [6, 7]. COSMIN emphasizes standards for PROMs based on classical measurement theory and may be less applicable to physical performance measures or those designed from other theoretical frameworks. Other tools used to assess outcome measure studies include a clinical measurement *Quality Appraisal for Clinical Measurement Research Reports Evaluation* [8] that assess study quality and a guide to critical appraisal of reliability studies that focuses on clinical application [9]. Different critical appraisal tools can lead to different conclusions [10, 11]. This is because they have different items and purposes.

Primary studies and SRs also require different critical tools since the key issues that indicate quality or bias are different. Measurement Tool to Assess systematic Reviews (AMSTAR 2) is the main tool we use to assess validity of SR and can be accessed at their online form link (https://amstar.ca/Amstar_Checklist.php). Strengths of AMSTAR 2 include simplicity, validated tools, and feasibility for novice evaluators. A downside of AMSTAR 2 is that some question descriptors are not tailored to SRs of OM and the value of a SR depends on the pool of available evidence, not just whether the SR was done correctly. However,

the intent of AMSTAR 2 questions can be adapted to assess SR in an OM context. Regardless of which critical appraisal tool is used to assess SRs, the decision on trust in the conclusions of an SR is an overall decision, not based on a cutoff score. COSMIN may be appropriate for some of types of reviews (PROM of classical test theory OM), but for theoretical and practical reasons we used AMSTAR 2.

A. Is it valid?

Strengths of the SR by Hao et al. [2] based on our AMSTAR 2 evaluation included a clear research (PICOT), duplicate screening and data extraction, use of a validated critical appraisal tool, the SR followed the COSMIN guideline for SRs of outcome measurement instruments, the protocol was registered at the International Prospective Register of Systematic Reviews (PROSPERO), and the authors had no conflicts of interest [2]. Although a narrative review was required due to the nature and volume of evidence, quality was considered when summarizing the evidence.

Partial quality was assigned in some areas as the rationale for the methods was not fully described. Specifically, excluded studies were not listed and some details of the search were not stated (e.g., grey literature searching and expert consultation). A meta-analysis was not performed.

The SR identified some limitations in the primary literature. Some studies' reliability was classified as doubtful because the tests and retests were administered in the same session. One study evaluated reliability, criterion validity, and hypotheses testing for construct validity, but was classified as inadequate because of a lack of proper statistical analysis to quantify the measurement properties. All other (7/9) studies were rated "adequate" to "very good" in the five domains assessed.

B. What are the results?

The SR by Hao et al., "*Performance-based outcome measures of upper extremity in virtual reality and telerehabilitation: a systematic review*," syn-

thesized 9 cross-sectional studies that enrolled a total of 210 patients with neurological conditions and 184 healthy subjects [2]. This is a relatively small pool of research, but not surprising given that only recently has remote clinical practice needed to focus on assessing patient performance. Seven studies included patients with stroke, one included patient with Parkinson's disease, and one included patient with a mix of neurological conditions, including cervical spinal cord injury, Guillain-Barré syndrome, and meningitis. Four studies evaluated the Box and Block Test in virtual reality, and three studies implemented the Fugl-Meyer Assessment in telerehabilitation. Single studies evaluated the Action Research Arm Test and Peg Insertion Test in virtual reality.

The authors stated that their findings supported the feasibility of implementing upper extremity PBOMs in virtual reality and telerehabilitation. Most studies demonstrated adequate reliability. The validity of virtual reality and telerehabilitation assessments compared to traditional in-person assessments was variable; 6/9 studies reported good correlations between the two. Modifications of original standardized OM protocols were used in multiple studies to adapt assessment to digital technologies and remote assessments. For the Box and Block test, 2 studies evaluated haptic-based virtual reality to simulate the movements of grasping and transferring blocks and 2 studies used virtual reality with hand-tracking motion capture systems. A peg test was evaluated with a computer-based haptic device, although the interactive nature of the test was compromised.

C. Can I apply the results to my practice?

Studies were eligible for this SR if they met all the following criteria: Patients with upper extremity impairments due to neurological disorders, PBOMs of upper extremity, including but not limited to the Fugl-Meyer Assessment, Box and Block Test, Action Research Arm Test, and Nine Hole Peg test, were assessed by virtual reality or telerehabilitation; measurement properties (reliability, validity, responsiveness) were assessed and reported.

When considering whether the results apply to practice, we must consider scope, quality, and feasibility. The evidence pool was small and recommendations might change as the volume and scope of evidence increases. The major issue is that the SR focused on neurological disorders so the results can not be generalized to musculoskeletal disorders, which form a substantial portion of telehealth practice. Findings from other patient populations might have improved the clinical relevance. However, the data can be applied with some confidence to patients with neurological conditions. The feasibility of the tests was underemphasized and technology availability and literacy in patients is a substantial issue in some technology solutions. As we move to measure motion, strength, coordination, and dexterity in upper extremity health problems, pragmatics and evidence will need to be considered. This is an important and emerging area of clinical research that is needed if we are going to ensure telehealth does not result in substandard, poorly monitored outcome assessment.

Resolution of the Clinical Scenario

The three components of evidence-based practice are evidence, clinical experience, and patient-values and preferences. There is weak evidence that sophisticated tools like virtual reality and haptic devices can take simple in-person dexterity tests like the Box and Block test and provide satisfactory measurement properties. However, the team decided that the processes studied were not clinically feasible for routine implementation as the tests depended on devices that were not amenable to routine clinical practice, nor were “patient friendly.” This field is rapid changing and artificial intelligence (AI) tools that can track movement are likely to be more feasible for clinical practice, but these have not yet appeared as methods for testing performance-based function tests for patients with neurological conditions, so the team decided to re-evaluate the decision in 6 months.

Appendix 1 Search Strategy and Results

Search Strategy

Search strategies are most comprehensive when they involve a librarian with expertise. For clinical practice having efficient methods to get to the best evidence quickly becomes equally important.

In 2009 COSMIN authors recommended a PubMed filter for studies on psychometric properties: #1 (construct search), #2 (population search), #3 (instrument search), #4 #1 (construct search) AND #2 (population search) AND #3 (instrument search) AND measurement properties filter, #5 (#4 NOT exclusion filter) [12]. Our strategy below is based on our experience finding OM papers but must be tailored based on the specific clinical research question or scenario.

Limitations: Abstract, Past 5 years, systematic reviews, Date November 30, 2024.

PubMed

Clinical queries are a great option for clinicians to rapidly access the most relevant high-quality research for specific types of clinical questions through the use of validated filters (e.g., treatment effectiveness, prognosis, diagnosis). Unfortunately, Clinical Queries does not have a filter for outcome measures. In general, when searching for psychometric evidence using the terms for psychometric properties (list separated by OR) AND (construct of interest/type of outcome measure) AND (outcome measure OR questionnaire OR test) AND (population OR region) will often provide an efficient and fairly accurate search. If searching for SRs this option can be selected in PubMed or added to the search.

(Reliability OR validity OR responsiveness OR Rasch OR content validity OR patient preference OR ceiling effects) AND (performance-based outcome measure OR range of motion OR strength) AND (upper extremity OR upper limb OR shoulder OR elbow OR wrist OR hand) AND (scoping review OR systematic review)

AND (telemedicine OR telehealth OR telerehabilitation OR virtual assessment OR remote care). We selected the last 5 years and having an abstract to filter out old evidence and editorial type papers. This revealed 7 papers—a manageable list. The most useful paper from this list was the article appraised- bolded below [2]. Reasons for not including the other 6 papers are also listed below.

Google Scholar

Google Scholar often performs well when looking for a very specific type of paper since you can use the specific terms to try and quickly focus on information you need. You can limit the search as we did to articles in a specific timeframe (we selected 2023+ since we know that we are interested in recent developments). A complication when searching for virtual OMs is that there are many different terms used in this field (telerehabilitation, telehealth, e-health, remote care, virtual care); and this can overlap with work on wearable sensors, virtual reality, etc.

Search terms: upper extremity, performance-based outcome measures, scoping review, systematic review, virtual care, remote care, telerehabilitation.

The paper selected for review that was highly relevant, and appeared in the top three listed by Google Scholar.

Using the search strategy above no relevant citations were retrieved from CINAHL.

In keeping with the need to be efficient in searching in clinical practice, we focused on complementary databases and search strategies that were more likely to produce focused results, which would have been a different strategy.

Retrieved (Duplications Removed); Highlighted Selected Paper

1. Fu Y, Zhang Y, Ye B, Babineau J, Zhao Y, Gao Z, Mihailidis A. smartphone- based hand function assessment: systematic review. *J Med Internet Res*. 2024 Sep 16;26:e51564.

doi: 10.2196/51564. PMID: 39283676; PMCID: PMC11443181.

Not selected: not performance based tests

2. Bortolani S, Brusa C, Rolle E, Monforte M, De Arcangelis V, Ricci E, Mongini TE, Tasca G. Technology outcome measures in neuromuscular disorders: a systematic review. *Eur J Neurol*. 2022 Apr;29(4):1266-1278. doi: 10.1111/ene.15235. Epub 2022 Jan 10. PMID: 34962693.

Not selected: not upper extremity population

3. Avila FR, Carter RE, McLeod CJ, Bruce CJ, Giardi D, Guliyeva G, Forte AJ. Accuracy of Wearable Sensor Technology in Hand Goniometry: a Systematic Review. *Hand (N Y)*. 2023 Mar;18(2):340-348. doi: 10.1177/15589447211014606. Epub 2021 May 25. PMID: 34032154; PMCID: PMC10035090.

Not selected: not performance based functional tests

4. **Hao J, Li Y, Remis A, He Z, Yao Z, Pu Y. Performance-based outcome measures of upper extremity in virtual reality and telerehabilitation: a systematic review. *Neurol Sci*. 2024 Mar;45(3):977-986. doi: 10.1007/s10072-023-07158-0. Epub 2023 Oct 31. PMID: 37904015.**

Selected

5. Zischke C, Simas V, Hing W, Milne N, Spittle A, Pope R. The utility of physiotherapy assessments delivered by telehealth: a systematic review. *J Glob Health*. 2021 Dec 18;11:04072. doi: 10.7189/jogh.11.04072. PMID: 34956637; PMCID: PMC8684795.

Not selected: not performance based tests

6. Terry D, Phan H, Peck B, Hills D, Kirschbaum M, Bishop J, Obamiro K, Hoang H, Nguyen H, Baker E, Schmitz D. Factors contributing to the recruitment and retention of rural pharmacist workforce: a systematic review. *BMC Health Serv Res*. 2021 Oct 5;21(1):1052. doi: 10.1186/s12913-021-07072-1. Erratum in: *BMC Health Serv Res*. 2021 Nov 16;21(1):1238. doi: 10.1186/s12913-021-07175-9. PMID: 34610827; PMCID: PMC8493699.

Not selected: not performance based tests; not clinical measurement

7. Zhang B, Fang Z, Nian K, Sun B, Ji B. The effects of telemedicine on Rotator cuff-related shoulder function and pain symptoms: a meta-analysis of randomized clinical trials. *J Orthop Surg Res.* 2024 Aug 14;19(1):478. doi:10.1186/s13018-024-04986-4. PMID: 39143625; PMCID: PMC11323625.

Not selected: not performance based tests clinical measurement

8. Shepherd J, Hansjee S, Divall P, Raval P, Singh HP. How do digital range of motion measurement devices 'measure-up' to traditional goniometry in assessing shoulder range of motion? A systematic review and meta-analysis. *Shoulder Elbow.* 2024 July;16(4):363-381. doi:10.1177/17585732231195554. Epub 2023 Aug 31. PMID: 39318409; PMCID: PMC11418675.

Not selected: not performance upper extremity functional tests

9. Hao J, Huang B, Remis A, He Z. The application of virtual reality to home- based rehabilitation for children and adolescents with cerebral palsy: a systematic review and meta-analysis. *Physiother Theory Pract.* 2024 July;40(7):1588-1608. doi:10.1080/09593985.2023.2184220. Epub 2023 Feb 27. PMID: 36847396.

Not selected: not clinical measurement, not the right population, interventional.

References

1. Melian C, Kieser D, Frampton C, Wyatt C, M. Teleconsultation in orthopaedic surgery: a systematic review and meta-analysis of patient and physician experiences. *J Telemed Telecare.* 2022;28(7):471-80. <https://doi.org/10.1177/1357633X20950995>.
2. Hao J, Li Y, Remis A, He Z, Yao Z, Pu Y. Performance-based outcome measures of upper extremity in virtual reality and telerehabilitation: a systematic review. *Neurol Sci.* 2024;45(3):977-86. <https://doi.org/10.1007/s10072-023-07158-0>.
3. Prinsen CAC, Mokkink LB, Bouter LM, Ionso J, Patrick DL, de Vet HCW, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res.* 2018;27(5):1147-57. <https://doi.org/10.1007/s11136-018-1798-3>.
4. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Qual Life Res.* 2021;30(8):2197-218. <https://doi.org/10.1007/s11136-021-02822-4>.
5. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, Terwee CB. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Qual Life Res.* 2018;27(5):1171-9. <https://doi.org/10.1007/s11136-017-1765-4>.
6. McKenna SP, Heaney A. Setting and maintaining standards for patient-reported outcome measures: can we rely on the COSMIN checklists? *J Med Econ.* 2021;24(1):502-11. <https://doi.org/10.1080/13696998.2021.1907092>.
7. Rosenkoetter U, Tate RL. Assessing features of psychometric assessment instruments: a comparison of the COSMIN checklist with other critical appraisal tools. *Brain Impair.* 2018;19(1):103-18. <https://doi.org/10.1017/BrImp.2017.29>.
8. McGee S, Sipsos T, Allin T, Chen C, Greco A, Bobos P, MacDermid J, CATWAD. Systematic review of the measurement properties of performance-based functional tests in patients with neck disorders. *BMJ Open.* 2019;9(11):e031242. <https://doi.org/10.1136/bmjopen-2019-031242>.
9. Bialocerkowski A, Klupp N, Bragge P. How to read and critically appraise a reliability article. *Int J Ther Rehabil.* 2010;17(3):114-20. <https://doi.org/10.12968/ijtr.2010.17.3.46743>.
10. Losilla JM, Oliveras I, Marin-Garcia JA, Vives J. Three risk of bias tools lead to opposite conclusions in observational research synthesis. *J Clin Epidemiol.* 2018;101:61-72. <https://doi.org/10.1016/j.jclinepi.2018.05.021>.
11. Banzi R, Cinquini M, Gonzalez-Lorenzo M, Pecoraro V, Capobussi M, Minozzi S. Quality assessment versus risk of bias in systematic reviews: AMSTAR and ROBIS had similar reliability but differed in their construct and applicability. *J Clin Epidemiol.* 2018;99:24-32. <https://doi.org/10.1016/j.jclinepi.2018.02.024>.
12. Terwee CB, Jansma EP, Riphagen II, De Vet HCW. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res.* 2009;18(8):1115-23. <https://doi.org/10.1007/s11136-009-9528-5>.

Patient-Reported Outcome Measures in Surgery: Development, Selection, and Evaluation

Manraj N. Kaur, Lucas Gallo, Anne F. Klassen,
and Andrea L. Pusic

Abbreviations

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CFA	Confirmatory Factor Analysis
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments
CTT	Classical Test Theory
EFA	Exploratory Factor Analysis
IRT	Item Response Theory
MID	Minimal Important Differences
PRO	Patient-reported outcome
PROM	Patient-Reported Outcome Measure
RMT	Rasch Measurement Theory
SDC	Smallest Detectable Change
SEM	Standard Error of Measurement

M. N. Kaur (✉) · A. L. Pusic
Patient-Reported Outcomes, Value & Experience (PROVE) Center, Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
e-mail: mkaur17@bwh.harvard.edu;
apusic@bwh.harvard.edu

L. Gallo
Division of Plastic and Reconstructive Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada
e-mail: lucas.gallo@medportal.ca

A. F. Klassen
Department of Pediatrics, McMaster University, Hamilton, ON, Canada
e-mail: aklass@mcmaster.ca

Clinical Scenario

You are a newly hired burn surgeon at a local trauma hospital. At the monthly burn research meeting, one of your senior colleagues suggested that the trauma team perform a randomized controlled trial comparing isolated steroid injection versus fractional CO₂ laser therapy with steroid injection for the treatment of hypertrophic flame burn scars on the extremity and body. Your team unanimously agrees that the primary outcome of this study should be the appearance and quality of a hypertrophic scar at 1 year following the onset of treatment. A colleague suggests that a 5-point Likert scale be used where a blinded surgeon evaluates the appearance of the scar at this time horizon. You feel that, since appearance is a patient-reported outcome (PRO), a patient reported outcome measure (PROM) would better reflect the patient's perspective on the effectiveness of these interventions. You inform your colleagues that you will review the literature to identify existing PROMs that may be used for this study and will present your recommendation at the next research meeting.

Literature Search

You begin your search on PubMed Medline where you include the key words "Burn" AND "Scar" AND "Patient-reported outcome measure" and limit the search to systematic reviews

performed in the last 10 years (January 2014–November 2024) (see Chap. 4). The search returns five articles (see Appendix). One article by Mundy et al. is a 2016 systematic review of existing PROMs developed and/or validated for patients with surgical or traumatic scars [1]. The authors of this review critically appraised these instruments based on the recommended steps set by the United States Food and Drug Administration for health-related outcome measures development and psychometric validation. Ultimately, the study concludes that all existing patient-reported scar measurement instruments demonstrated significant limitations and suggested that a future scar-focused PRO instrument be developed. Of the remaining articles, one is a systematic review by Carriere et al., which tabulates the frequency of various clinician-reported and patient reported outcome measurement tools to evaluate scars; however, this provides you with no additional information [2]. The remaining articles have limited relevance to your research question [3–5].

After reviewing the manuscript by Mundy et al. [1], you identify a follow-up study published by the same group of authors in 2018, titled “Development of a New Patient-reported Outcome Instrument to Evaluate Treatments for Scars: The SCAR-Q” [6]. You are encouraged that the SCAR-Q may be a suitable PROM for your proposed study. Given this, you review other articles to learn more about the importance of PROMs as well as the steps required to develop and select a PROM.

Introduction

It is much more important to know what sort of a patient has a disease, than what sort of a disease a patient has. (Sir William Osler)

The field of surgery has had a long history of focusing on “objective” outcomes or observable aspects of disease or condition (e.g., mortality, graft patency rate, tumor margin status, and surgical site infection), where patient was viewed as the passive recipient. While surgeon-reported

outcomes remain vital in the evaluation of treatment effectiveness, in the past 50 years, there has been an increasing recognition that patient-reported outcomes are critical to planning, organizing, delivering, and reimbursing patient-centered health care [7]. Several examples of the application of patient-reported outcome data in health care exist in the literature. Such data has been used to enhance clinician-patient communication (set expectations presurgery, empower patients to engage in surgical decision making), to aid in patient monitoring (early identification and management of adverse events), to support surgical care planning (alignment of care with patient’s goals, preferences, and values), for peer benchmarking (support comparison of outcomes across providers in a health system or jurisdiction), for quality improvement (evaluate treatment effectiveness, establish standards of care), and to provide comprehensive and tailored care (integration of social needs, support discharge planning) [8–11].

Patient-reported outcomes (PROs) are any report of the status of a patient’s health condition that comes directly from the patient, without interpretation of the patient’s response by a clinician or anyone else [12]. Put simply, PROs are outcomes that can only be known by asking the patient directly and not through observation, physical examination, diagnostic, or laboratory investigations. Some examples of PROs include pain, anxiety, social well-being, body image, satisfaction with appearance, health-related quality of life, satisfaction with preoperative information, and satisfaction with one’s experience of health care. A *PRO measure (PROM, for short)* is a questionnaire that is designed to measure the treatment outcome (benefit or risk), health status, disease, or health condition from the patient perspective. The components of a PROM are shown in Fig. 8.1, and key definitions are included in the *Abbreviations*.

For a PROM to be helpful in clinical care, research, and quality improvement, it must measure the PRO (“construct”) of interest reliably and validly in the target population and clinical context. The PROM should also be responsive to

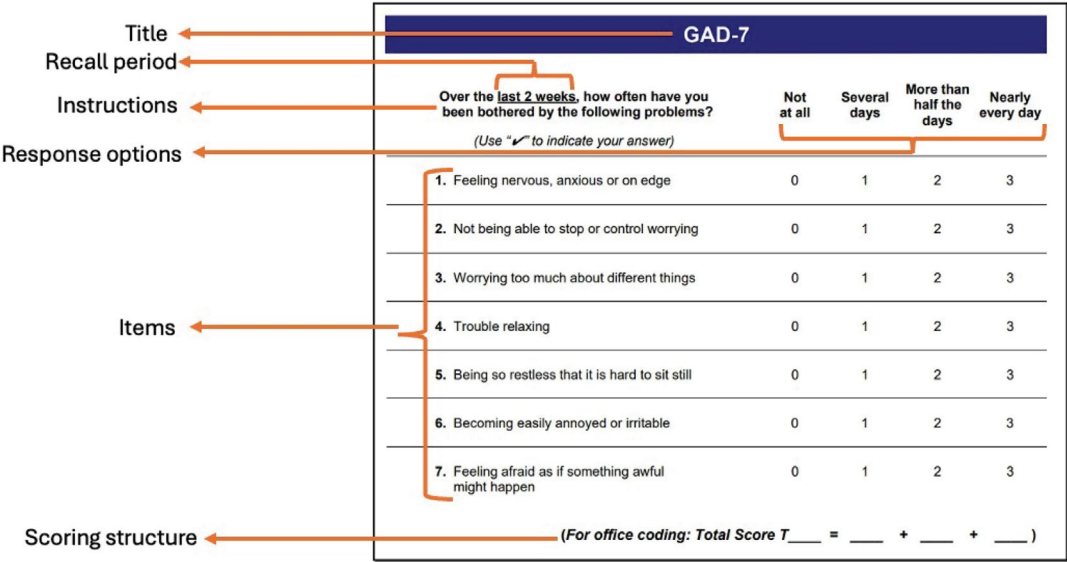


Fig. 8.1 Components of a PROM using the Generalized Anxiety Disorder-7 (GAD-7) as an example

Table 8.1 Recommended expertise required to develop a PROM

Qualitative researcher.
Psychometrician/quantitative psychologist.
Quantitative researcher.
Mixed methods researcher.
Survey designer.
Statistician.
Individuals with lived experience of the construct (patients).
Clinician(s) experienced in working with target population.
Community members experienced in/working with target population.
End user(s) of the PROM (e.g., clinicians, hospital administrators).

measuring change that may occur in the PRO because of intervention or the natural progression of a health condition. Furthermore, the intended purpose of the PROM should be defined, that is, whether one wants to use the PROM as a primary, key secondary, or exploratory endpoint [12]. The following sections will describe methods to develop a PROM, measurement theory associated with PROMs, and key measurement properties to consider when deciding whether to select a specific PROM.

How to Develop a PROM

PROM development is an iterative activity that can take several years depending on its scope. Developing and validating a PROM requires multidisciplinary expertise (Table 8.1) and systematic involvement of stakeholders, including patients and clinicians.

Broadly, the development of a PROM can be conceptualized in five components: (1) define the construct of interest, target population, context of use; (2) develop a conceptual framework and related measurement model; (3) develop and select items for the first draft of the PROM; (4) pilot-test the PROM; and (5) field-test the PROM to establish its psychometric properties and scoring function. These components are described in detail below.

Defining the Construct of Interest, Target Population, and Context of Use

The first question one should ask when developing a PROM is, “What construct(s) do we want to measure?” A construct is defined as a “well-defined and precisely demarcated subject of mea-

surement” [13]. This term is also defined as the unobservable or latent trait or characteristic that needs to be quantified. Some examples of constructs include pain, fatigue, anxiety, or depression. Once the construct is identified, the second question is “Who do I want to measure the construct in?” Defining the target population is critical because the questions that are suitable for adults may not be appropriate for pediatric population, and the questions may not work the same way across different culture or language subgroups. Similarly, awareness of the health literacy and reading literacy of the target population will drive the comprehensibility of the questions.

Develop a Conceptual Framework

The relationship between the construct and the items is known as the conceptual framework. The conceptual framework is a central element of a PROM because it determines the measurement theory to be used in the development and evaluation of the PROM [13]. There are two main types of conceptual frameworks—reflective and formative. A reflective framework is when the construct manifests itself in the items. The items in a reflective framework are called “effect indicators.” A formative framework differs in that the items cause a construct and as such are called “causal” indicators. The easiest way to know if the PROM is based on a formative or reflective framework is to ask, “will the responses to the items change when the construct changes?” If the answer is yes, then you have a reflective framework. If the responses to the items do not align with changes in the construct, then the framework is formative. It is important to note that some complex constructs, such as quality of life, may have both reflective and formative components [13]. Figure 8.2 shows a generic example of a conceptual framework of PROM.

The conceptual framework is straightforward when the concept(s) of interest can be measured by a single item(s) (e.g., pain intensity [14], self-reported mental health [15]). However, single items often fail to provide full coverage of a single

concept of interest or domain (also called subconcept; e.g., physical function) and multiple items are needed for adequate measurement. Multidomain conceptual frameworks are possible (e.g., Multidimensional Fatigue Inventory [16]) and involve multiple items that map out the multiple domains. Single-item PROMs and multidomain PROMs are generally not ideal for use in clinical trials and clinical care. Single-item PROMs do not have the reliability and validity and rarely provide enough evidence to support the concept of interest [17]. Multidomain PROMs that produce a single score are also problematic because the items that make up the domains may change in different directions (some better, some worse) or some change while others remain static, making it hard to interpret the scores [18, 19].

The conceptual framework evolves over the course of the development of a PROM and qualitative methods are used to confirm the comprehensiveness of the conceptual framework. The final conceptual framework should be confirmed by examination of the measurement properties of the PROM. Conceptual framework can be developed with concepts identified in the literature and with qualitative interviews or focus groups with patients and other relevant stakeholders (e.g., clinicians, nurses). Since PROMs capture what is relevant to the patients, the inclusion of the patient perspective is paramount to the development of a conceptual framework.

One-on-one interviews and focus groups are the two most common qualitative methods used to develop a conceptual framework for a PROM. While no one method is ideal, the choice of qualitative method depends on the context, target population, and resource availability and should be documented. Before selecting a qualitative data collection method, it is important to consider the potential strengths and limitations of the methods [20]. The PROM development guidance does not specify a set number of interviews or focus groups to be conducted; rather the emphasis is placed on whether the concept has been mapped out comprehensively and saturation has been achieved.

Several papers have established good practices for conducting qualitative interviews.

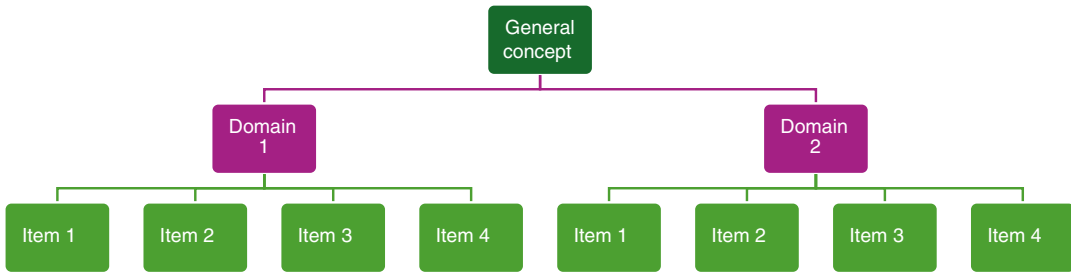


Fig. 8.2 Example of a conceptual framework of a PROM

[21–26] For one-on-one interviews and focus groups, an interview guide is drafted based on the literature and/or expert input. The structure of the interviews can be semi-structured or non-structured. An experienced/trained qualitative interviewer should conduct the interviews. Generally, at the beginning of an interview, the interviewer starts with open-ended questions (e.g., can you tell me when you were diagnosed?) and then moves on to more targeted questions about the concept of interest, while still keeping the questions open-ended to avoid yes/no responses (e.g., you mentioned you struggled with fatigue during chemotherapy, can you tell me more about that?...). For an example of an interview guide used to develop a PROM, see the protocol for the development of the GENDER-Q, a PROM for adults seeking gender-affirming care [27]. For pediatric patients or patients with limited literacy, qualitative methods that art-based activities have recently become popular [28–30]. For rare conditions, social media has been used to elicit relevant concepts important to patients for PROM development [31]. The last stage of the interview should include closing questions that invite participants to mention anything that was relevant to them that was missed or incompletely discussed.

A data analysis plan and use of qualitative data analysis software are established during the planning stages of the study to develop the conceptual framework. Commonly used analytic methods for PROM development studies include content and thematic analysis. While both are qualitative descriptive approaches and employ a low level of interpretation in comparison to other qualitative methods, such as phenomenology or

grounded theory, there are subtle differences between the two that have been elaborated in by Vaismoradi et al. [32], Humble and Mozelius [33], and a blended approach has been proposed by Neuendorf [34]. For a summary of key differences, see Table 8.2. Depending on the concept and available literature, coding can be inductive, deductive, or a combination of inductive and deductive approaches. A codebook should ideally be designed prior to coding the interviews and refined iteratively over the course of qualitative work. A codebook allows one to consistently code the qualitative data and enhances trustworthiness of the analysis. The steps to test and create the codebook are shown in Fig. 8.3.

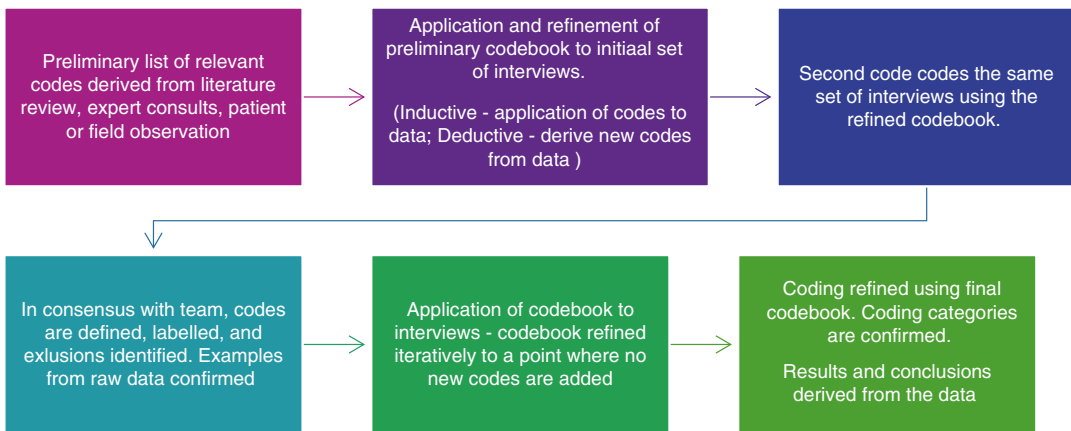
Once the data are coded, the outcome will be a conceptual framework consisting of items, domains, and their interconnections, along with an item pool. An item pool is the list of patient quotes that exemplify the concepts of interest mentioned by the interviewees. The item pool is then used to develop the first draft of the PROM.

Develop and Select Items for First Draft of the PROM

Best practices for item development have been suggested in the literature [19, 37]. The items should be developed using the words used by patients such that the language for each item seems natural and familiar (i.e., resonates). The items should also be evaluated for translatability, particularly if the objective is to develop a PROM with global applicability. Additionally, the items must be assessed for their appropriateness in relation to the literacy, health literacy, and numer-

Table 8.2 Key differences between content analysis and thematic analysis [32]

	Content analysis	Thematic analysis
Goal	Objective, quantitative analysis of message characteristic	Develop a story from the text
Purpose	Describe characteristics of a document's content by examining who says what, to whom, and with what effect [35]	To identify, analyze, and report patterns (themes) within data [36]
Philosophical background	Communication theory, factist perspective	Realist/essentialist and constructivist, factist perspective
Procedure	Emphasizes content and integration of latent and manifest contents	Division of manifest and latent contents, danger of missing context and possibility of finding a theme based on frequency of occurrence
Rigor	Peer-checking	Peer-checking not always possible

**Fig. 8.3** Steps to create and test a qualitative codebook

acy capabilities of the target population. Use of incomplete, leading, poorly worded items should be avoided. Items should avoid double-barreled or multibarreled questions (e.g., how often do you experience pain and how intense is it?) as well as the use of double negatives (e.g., how often do you not engage in exercise when you have no pain?). Terms that have multiple meanings should be avoided (e.g., fair, pretty good, not bad, just) and items should be specific such that there is no ambiguity in the response (e.g., item asking about pain can be about current pain, worst pain, average pain in last 24 h, or average pain last week). The instructions should be clear and specify the recall period. The most common recall periods for PROMs include 24 h, and 1 or 2 weeks. The response options should be derived from the qualitative data and reflect how inter-

viewees described the concept [e.g., was fatigue described in terms of intensity (mild, moderate, severe), frequency (none of the time, some of the time, all of the time), interference with daily activities (not at all, a little, quite a bit, all the time)]. Different types of response options include checklists, binary responses (yes/no), rankings, rating scales, and visual analogue scale [38]. All the response options have their strengths and limitations in particular context of use. Response options with many categories, or unlabeled categories, tend to be of limited value. To avoid such problems, developers are advised to create simple questions, 4–5 response options, and ensure that all response options are labeled.

The preliminary draft of the PROM may include more than one item that is measuring the same concept and more than one response option

choice, to be then evaluated in a pilot test. This approach provides the developer with the opportunity to pick a subset of items to include in the final instrument based on their psychometric performance.

Demonstrate Content Validity of the Draft PROM

When the initial draft of the PROM is prepared, feedback on components of the PROM should be sought from representative sample of patients (ideal), clinician experts, and other relevant stakeholders. The evaluation should include documentation that the concepts represented in the PROM's conceptual framework are confirmed, and that the items are easy to understand and relevant to the target population [12]. This measurement property is known as content validity, and it is the most important measurement property of a PROM [39]. Evidence of reliability and construct validity are irrelevant if content validity is not demonstrated before the PROM is field tested.

Content validity is the extent to which the PROM measures the concept of interest. Content validity is ideally supported by qualitative evidence that demonstrates that the items and domains of a PROM are easy to understand, relevant, and comprehensive relative to its intended measurement concept, population, and use [12]. Content validity is specific to the population, condition or treatment, and context to be studied. This measurement property is usually examined with cognitive debriefing interviews with population of interest using the "Think Aloud" [40] technique whereby the respondent is asked to review the scale title, instructions, recall period, response options, and the items. A survey in a larger sample of participants can also be used to establish content validity. The participant is also asked to provide feedback to determine if the components of the PROM are easy to understand, relevant to their experience and if the PROM is comprehensive. For sample questions for cognitive debriefing interviews, see evaluation of PROMIS Fatigue items by Christodoulou et al. [41]. Where relevant (for PROMs based on reflec-

tive models), participants are asked if their responses on an item will change if the underlying measurement concept changes and to point out missing items. The cognitive debriefing interviews are conducted in rounds (e.g., 3–5 participants or stakeholders) and the PROM is refined between rounds. Items can be added, revised, or dropped for reasons including clarity or relevance, response range, lack of variability (all respondents give same answer), variability in response without any logical explanation, inability to detect change, redundancy or issues with recall period [12, 19, 42]. The COSMIN guidance for content validity recommends conducting a minimum of 7 cognitive debriefing interviews per scale or PROM or collecting quantitative data on content validity from at least 50 respondents using a survey approach. Fundamentally, pilot testing is a thorough qualitative analysis of the PROM in a small representative sample of the target population. This testing is followed by evaluation of the PROM in a large, diverse representative sample of the target population in a field test.

Field-Test the PROM

The aim of the field test is to reduce the number of items and examine the dimensionality of a PROM in a large representative sample of target population. Single-item instruments do not need to be item-reduced, and they measure only one domain; hence, the field test step is not necessary, and they go from Step—Demonstrate content validity of the draft PROM to the examination of their measurement properties of reliability, validity, and responsiveness. For the field test, the data should be collected from a few hundred patients ($\geq 250/\text{PROM}$) [43]. Ideally, the sample should be diverse in key relevant characteristics of the sample (e.g., age, sex, gender, education attainment). A measurement theory, that is, a theory about how scores generated by items represent the construct to be measured, is chosen. Notably, only PROMs about latent (unobservable) constructs require a measurement theory. The notable theories in the field of health measurement

derived from psychoeducational testing include Classical Test Theory (CTT), Rasch Measurement Theory (RMT), and Item Response Theory (IRT). RMT and IRT are modern psychometric theories that address limitations of the CTT. Below, we will briefly review key characteristics of these measurement theories followed by examination of field test data.

Measurement Theories

Classical Test Theory (CTT)

Classical Test Theory (CTT) posits that a PROM score consists of a “true” score (T) representing the actual level of the attribute or trait of interest, and a random error (E). This concept is expressed mathematically as $X = T + E$, where X is the observed score [44]. In CTT, the true score reflects the expected score a person would achieve across infinite, error-free measurements. However, only the observed score X is available in practice. True scores are intended to capture the real level of an individual’s attribute, construct, or trait that the PROM is designed to measure. Assuming item responses are coded to reflect increasing levels of the attribute, a higher true score should correspond to higher item responses on the underlying construct.

In CTT, errors are assumed to be random and do not systematically affect scores—errors have a mean of zero, meaning they can increase or decrease scores in equal measure. This implies that, over multiple hypothetical measurements, the errors would average out to zero. Additionally, in CTT, random errors are assumed to be uncorrelated with true scores, ensuring no systematic bias in how error affects different individuals’ scores. CTT also assumes a linear relationship between observed and true scores, so observed scores directly reflect true scores plus random error without non-linear interactions.

CTT’s strictest assumptions about items being “parallel” imply that each item measures the same true score with identical true score variances, equal error variances, and unidimensionality, meaning each item contributes equally to the

scale’s reliability. Furthermore, CTT assumes that errors are uncorrelated across items and with true scores, so error on one item is unrelated to error on another, maintaining the uniqueness of each item’s error.

Reliability in CTT is assessed by examining variance shared among items, with true score variance ideally comprising a substantial portion of the total variance. High item-item or item-total correlations suggest alignment with the underlying construct. Thus, in CTT, adding items tends to cancel out cumulative error and improve reliability, similar to how larger sample sizes in population studies yield more accurate estimates.

Advantages of CTT:

1. Familiarity: Many clinicians, including surgeons, are familiar with basic CTT concepts, having encountered these principles in measurement theory.
2. Widespread use: Most scales prevalent in the clinical and surgical literature are based on CTT principles, and CTT analysis methods are widely accessible in statistical software, which typically include programs for factor analysis and reliability estimation (e.g., Cronbach’s alpha).
3. Applicability to reflective models: CTT often suits PROMs that follow reflective models, commonly used in social sciences.
4. Practicality with item quality: Individual items do not need to be optimized in CTT; even items that relate modestly to the construct can still contribute to reliability. By increasing the number of items, it is possible to achieve almost any desired level of reliability.

Disadvantages of CTT:

1. Redundancy in measurement: Because CTT-based scales can be lengthy and contain highly similar items, they may result in redundant measurement, lacking the scrutiny of item quality seen in other frameworks.
2. Sensitivity limits: CTT scales may display differential sensitivity across the score range, being more responsive to changes in the middle than at the extremes. A 2-point difference near the center might reflect a smaller true

score difference than the same spread at the high or low ends.

3. **Sample dependency:** Parameter estimates in CTT are dependent on the specific sample studied, as item characteristics and scale reliability are calculated based on correlations within that sample. This dependency can complicate data interpretation across different samples, as scales may not yield comparable results when sample characteristics vary.

In sum, while CTT provides an accessible and widely used framework for scale development, its limitations underscore the importance of careful sample selection and awareness of scale sensitivity across the score range. For an in-depth review of CTT, refer to [44–48].

Modern Psychometric Theories—RMT and IRT

Unlike Classical Test Theory (CTT), which focuses on observed scores, IRT examines how specific items function across different levels of the underlying trait being measured. IRT posits that the probability of a correct response to an item is a function of both the individual's ability (or trait level) and the characteristics of the item itself, such as its difficulty and discrimination parameters. In IRT, each item is associated with a characteristic curve that represents the likelihood of a respondent with a particular ability level answering the item correctly. This allows for a nuanced understanding of how individual differences influence item responses, enabling more tailored and accurate assessments. RMT, a specific form of IRT, emphasizes the requirement for items to function consistently across different populations. RMT model ensure that the measurements are invariant, meaning the estimated abilities and item difficulties remain constant regardless of the sample.

Advantages of modern psychometric theories:

1. **Individualized Measurement:** IRT and RMT allows for a detailed analysis of how different individuals interact with items, facilitating the development of tailored assessments that can better capture a respondent's abilities or traits.
2. **Item-Level Insights:** IRT provides information on item characteristics, including difficulty and discrimination, allowing researchers to refine instruments by selecting or modifying items based on their performance across populations.
3. **Ability Estimation:** IRT produces more accurate estimates of an individual's latent trait level, as it considers the specific item responses, which can be particularly beneficial for adaptive testing environments.
4. **Measurement Invariance:** RMT ensures that item difficulties and person abilities are not sample-dependent, which enhances the comparability of scores across different groups or populations.
5. **Assumptions of Unidimensionality:** Both IRT and RMT assume that the items measure a single latent trait, which may not always hold true in practice, especially in multidimensional constructs; however, allows for reduced clinician and patient burden—both logistic and cognitive—since they are only responding to items related to one construct and one construct only.
6. **Efficiency:** IRT can improve the efficiency of assessments, allowing for shorter forms to be created without sacrificing measurement quality by selecting items that maximize information based on a respondent's ability level.

Disadvantages of modern psychometric theories:

1. **Complexity of implementation:** IRT and RMT require a more advanced understanding of statistics and programming, making them less accessible to researchers and surgeons unfamiliar with these methodologies.
2. **Sample size requirements:** These approaches typically require larger sample sizes to produce stable estimates for item parameters, which can pose challenges in studies with limited participant availability.
3. **Sensitivity to model fit:** The appropriateness of the IRT or RMT model is contingent upon the fit of the data to the model assumptions. Poor model fit can lead to inaccurate interpretations and invalid conclusions.

4. Limited literature: While gaining traction, the use of IRT and Rasch methods may still be less prevalent than CTT in certain fields, leading to a potential lack of familiarity and resources for surgeons seeking to implement these approaches.

In summary, modern psychometric theories, particularly IRT and Rasch measurement, offer sophisticated tools for analyzing item responses and estimating latent traits. Their strengths lie in their ability to provide individualized, invariant, and efficient measurements. However, these methods also present challenges related to complexity, sample requirements, and model assumptions, necessitating careful consideration when choosing a measurement approach.

For an in-depth review of modern psychometric theories, refer [49–53]. For a worked example comparing CTT, RMT, and IRT, see Ref. [54].

Steps Taken to Analyze Field Test Data

Examine missing data: The first step is to examine item scores and assess missing scores and patterns of missing score data. If scores are missing consistently for some items, then a closer look at the content and phrasing of the item is warranted as it may mean that the respondent did not understand the item, the items were not applicable to them, or the response options provided were confusing or inadequate. Generally, less than 3% missing scores are acceptable while more than 15% are not [13].

Distribution of item scores: This step ensures that all response option levels are informative and allows for the identification of items for which a large proportion of the sample has the same score. It is important to check the category frequency distribution for each response option, particularly focusing on the extreme response options. In modern psychometric theory analysis, it is also possible to plot item characteristic curves which display the probability of endorsing a particular response option for each item as a function of the underlying trait being measured. When the thresholds are disordered, the resulting

curves will not show the expected monotonic increase. Instead, they may cross over or appear irregular, indicating that the ordering of the response categories is not consistent or logical. If response options are disordered, one should consider reducing the number of response options by merging adjacent categories. The item characteristic curves also provide an indicator of the discriminative power and difficulty of an item.

Examine items in the scales: Examining items in scales is an iterative process in which both the items and their response options are assessed. This process involves testing items against the assumptions of the underlying measurement theory, allowing for the removal or reintroduction of items until a satisfactory solution is achieved.

- Using the CTT approach—factor analysis is the most employed method for examining the dimensionality of data and reducing the number of items in a scale. This approach represents a statistically derived means to identify the items to include in scales as opposed to a qualitative concept-driven approach used in IRT. The fundamental principle is that highly correlated items will cluster together within a single factor, enabling the deletion of items with low or no correlations. There are two main types of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA is typically applied during the development of a PROM, while CFA is primarily used to assess construct validity. In EFA, the process begins with the examination of an inter-item correlation matrix, identifying items that exhibit low correlations (below 0.2), high correlations (above 0.9), or negative correlations. Relevant factors are then extracted based on eigenvalues, retaining those greater than 1, and/or by assessing the relative contribution of each additional factor, often visualized through a scree plot. Rotation methods, particularly orthogonal rotation techniques like varimax, can be applied to enhance the interpretability of the factors. The factors are then labeled appropriately to indicate the constructs being measured. For scoring, CTT relies on the total score for a scale

and assumes equal contributions from all items [13, 47].

- Using the modern psychometric approach — IRT and RMT encompass several models tailored to different types of data and measurement scenarios. Common IRT models include the 1-parameter logistic model (1PL), which is synonymous with RMT and focuses solely on item difficulty; the 2-parameter model (2PL), which incorporates both item difficulty and discrimination; and the 3-parameter model (3PL), which adds a guessing parameter to account for the likelihood of a correct response by chance. RMT looks to see how data collected for a set of items hypothesized to form a scale fits the Rasch model (as opposed to the CTT approach that allows factor analysis to determine which items form which scales). This approach emphasizes the assumption of equal intervals between response categories (i.e., it operates on an interval scale) and is particularly valued for its strict adherence to unidimensionality. When assessing data fit to the Rasch model a range of tests are performed to examine item and scale performance. Items that fit the Rasch model have nonsignificant Chi-square values after Bonferroni adjustment and fit residuals that are ideally within the range of -2.5 to $+2.5$. In contrast, IRT model fit can be examined using likelihood ratio tests and information criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). In terms of scoring, RMT provides a probabilistic scoring approach based on item difficulty and individual ability, emphasizing unidimensionality; whereas IRT offers multiple models that capture the complexity of item response patterns, incorporating factors like discrimination and guessing to enhance scoring precision [49–53].

Examine internal consistency: Internal consistency, as defined by COSMIN, refers to the degree of interrelatedness among items within a scale [55, 56]. In the context of FA, examining internal consistency is also essential for item reduction as items highly correlated items (>0.7)

are considered redundant and could be considered for deletion. This assessment involves evaluating both inter-item correlations and item-total correlations. The most well-known statistic for measuring internal consistency is Cronbach's alpha, which represents the mean of the correlations obtained by splitting the PROM into two halves and assessing the correlation between the two resulting scores. As per COSMIN, internal consistency is considered an aspect of “reliability” of an instrument.

Examine measurement properties: Per the COSMIN taxonomy of measurement properties, the three main properties are: reliability, validity and responsiveness (see Fig. 8.4). The COSMIN definitions of these measurement properties and aspects of the measurement properties are provided in Table 8.3.

Examine floor and ceiling effects: Floor and ceiling effects can occur if more than 15% of the respondents achieve the lowest or highest possible score, respectively [57]. Floor and ceiling effects generally occur when the respondent population is more or less diseased than the population for which the PROM was originally developed. This effect has consequences for responsiveness and interpretability of an instrument.

Interpret scores: Scores can be interpreted using normative scores from general population or scores from well-known groups. Change scores (e.g., change from pre- to postoperative scores) can be interpreted by characterizing the relationship with change scores on other known instruments, using smallest detectable change, or minimal important differences (MIDs). The smallest detectable change (SDC) is calculated as $SDC = SEM \times z$, where SEM is the Standard Error of Measurement, which quantifies the amount of error in an individual score and is calculated as $SEM = SD \times \text{sq. root of } (1 - \text{Reliability})$ (SD is the standard deviation of the PROM scores, and Reliability is the reliability coefficient of the measurement tool), and z is the z -score that corresponds to the desired confidence level. For example, for a 95% confidence level, z is approximately 1.96. For a 90% confidence level, z is approximately 1.645. An MID can be estimated using anchor-based or distribution-based methods [58–60].

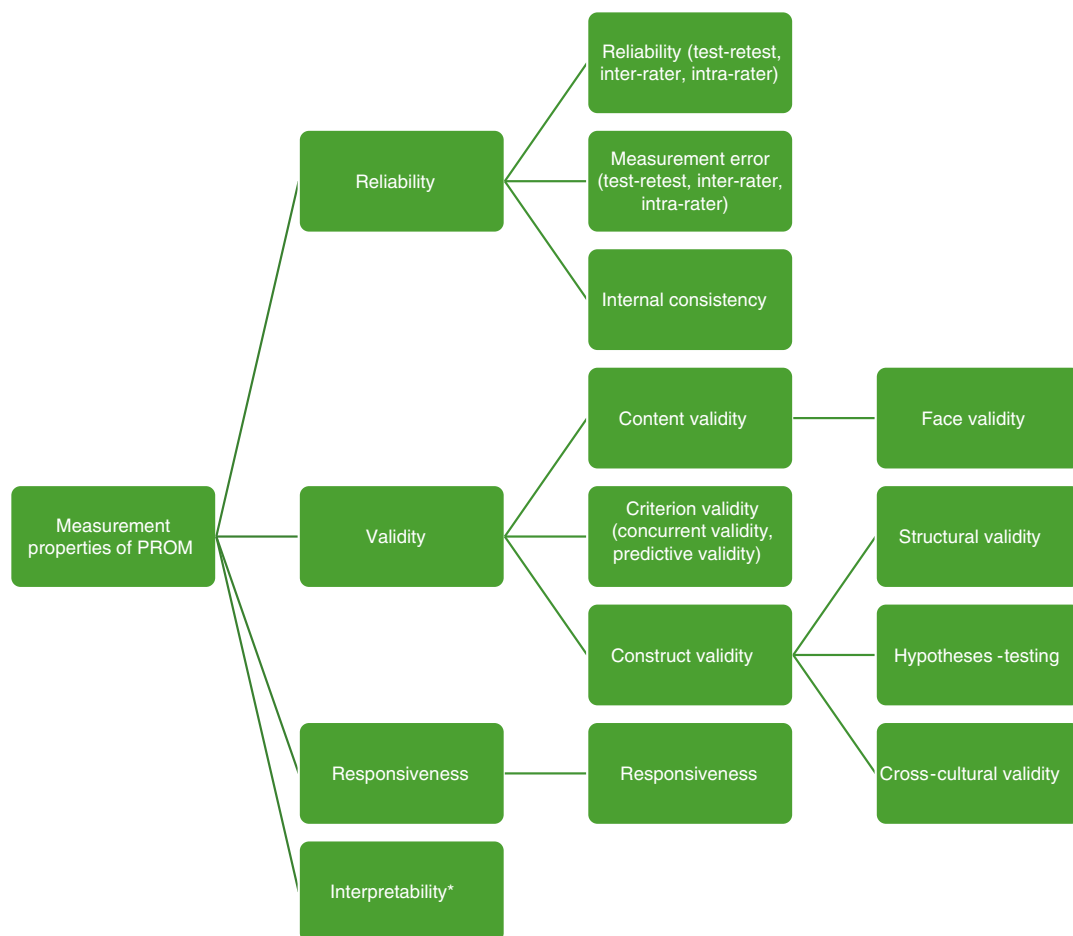


Fig. 8.4 Measurement properties of PROMs. (Adapted from COSMIN)

Table 8.3 COSMIN definition of measurement properties and aspects of measurement properties (Reproduced from F Ref. [61])

Term Domain measurement aspect of a property measurement property			Definition
Reliability			The degree to which the measurement is free from measurement error
Reliability			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: For example, using different sets of items from the same PROM (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons (raters or responders) on different occasions (intra-rater)
	Internal consistency		The degree of the interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements due to true differences between patients
	Measurement error		The systematic and random error of a respondent's score that is not attributed to true changes in the construct to be measured
Validity			The degree to which a PROM measures the construct(s) it purports to measure
	Content validity		The degree to which the content of a PROM is an adequate reflection of the construct to be measured
		Face validity	The degree to which (the items of) a PROM looks as though they are an adequate reflection of the construct to be measured
	Construct validity		The degree to which the scores of a PROM are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the PROM validly measures the construct to be measured
		Structural validity	The degree to which the scores of a PROM are an adequate reflection of the dimensionality of the construct to be measured
		Hypotheses testing	The degree to which the scores of a PROM are consistent with hypotheses (for instance, with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the PROM validly measures the construct to be measured (identical to construct validity)
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted PROM are an adequate reflection of the performance of the items of the original version of the PROM
	Criterion validity		The degree to which the scores of a PROM are an adequate reflection of a gold standard
Responsiveness			The ability of a PROM to detect change over time in the construct to be measured
Interpretability ^a			The degree to which one can assign qualitative meaning—That is, clinical or commonly understood connotations—To a PROM's quantitative score or change in scores

^aInterpretability is not a measurement property of a PROM, but rather an important characteristic of a PROM

How to Select a PROM

Developing a PROM is a time- and resource-intensive endeavor and requires multidisciplinary expertise and extensive patient input. Consequently, the first step that developers should take before creating a new PROM is to conduct a systematic review of the literature to identify existing PROMs and their measurement properties. If any PROMs are found for the topic of interest, the next step is to evaluate if the PROMs seem appropriate based on their content and if they have been validated in the population and context of use of interest. These steps are beneficial for the following reasons: (1) it prevents one from adding a new PROM to the literature (and contribute to research waste) when one that measures the construct of interest in the targeted population already exists; (2) the available resources can be directed to validating, adapting, and translating the existing PROM to the relevant context of use; (3) using the existing PROM allows for comparison of results across different studies; and at the very least (4) it gives one an idea of the concepts that have been measured in the literature for the topic of interest. Several types of systematic reviews of PROMs and their measurement properties exist in the literature [62]. One may do a systematic review to identify and assess all available studies on measurement properties of:

1. A single PROM of interest.
2. Commonly used PROMs that assess the construct of interest in the target population.
3. All PROMs that assess the construct of interest in the target population.
4. All PROMs that assess the construct of interest irrespective of the target population.
5. All PROMs used in the target population irrespective of the construct.

All five types of reviews may be restricted to “validated” PROMs to target the search, especially for clinical areas where a lot of PROMs exist (e.g., oncology) making the review particularly useful for selecting a PROM. Several guidance documents are available in the litera-

ture on how to conduct a systematic review of the literature, but the one that is most used and developed specifically for assessing measurement properties of PROMs was developed by the COSMIN (Consensus-based Standards for the selection of health Measurement INstruments) initiative [61]. The COSMIN initiative has developed a consecutive 10-step procedure to guide systematic reviews specifically for PROMs that can be broadly organized into 3-key steps. The first step is to conduct the literature search and identify relevant studies. Step 2 is to evaluate the quality of the eligible studies, the measurement properties and acceptability of the PROM. Step 3 is to formulate the recommendations and reporting of the systematic review results. A sensitive search filter for finding studies on measurement properties, risk of bias checklist for assessing the quality of studies and templates for data extraction and reporting are available at no charge from the cosmin.nl website (<https://www.cosmin.nl/tools/guideline-conducting-systematic-review-outcome-measures/>). The searchable COSMIN Database of Systematic Reviews contains 2181 reviews (as of October 25, 2024) and may be a helpful resource for surgeons planning to conduct a systematic review to identify the best PROM for their intended purpose. If an existing PROM that measures the construct of interest in the target population with established measurement properties does not exist, the development of a new PROM may be warranted.

If a PROM that adequately measures the construct of interest exists but has not been validated in the target population, the content validity of the PROM should be established and its measurement properties examined (see above: 4. Demonstrate content validity of the draft PROM).

If a validated PROM that adequately measures the construct of interest exists and has been validated in the target population, the user may consider the characteristics of the PROM that are important for acceptability and implementation of PROM data collection.

Applying the Literature to Patients

How Do You Interpret Results?

Being able to assign qualitative meaning to a quantitative score from a PROM or interpretability is of paramount importance. If a measurement tool shows validity and reliability but has not measured the issues that matter to patients, it is not useful [63]. Calculating the minimal important difference (MID) and standard error of measurement (SEM) is one approach to establishing interpretability of a PROM. Defined by Jaeschke and colleagues, MID refers to “the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate in the absence of troublesome side effects and excessive cost, a change in the patient’s management [64]. The MID should be defined to provide information regarding what change in score would be considered to be clinically meaningful [63, 64]. For an in-depth understanding of MID, refer to [58–60].

Critical Appraisal of the SCAR-Q

After reviewing the published literature regarding the development (Steps 1–4) and field test of the SCAR-Q (Step 5) [6, 65], you ask yourself three questions: (1) Is the PROM valid?, (2) Is the PROM reliable?, and (3) Is the PROM responsive to change?

Is the PROM Valid?

Following a review of the development of the SCAR-Q (Steps 1–4), you determine that this PROM was created through in-depth qualitative interviews followed by rigorous cognitive debriefing interviews with patients of varying ages, scar etiologies (i.e., including burn etiologies), and scar locations, thereby establishing the content validity of this PROM for a diverse range of adult and pediatric patients with surgical, traumatic, and burn scars. Notably, the authors of the SCAR-Q reference

and adhere to the recommended methods for PROM development [12, 19, 39, 42, 55]. You determine that a trained qualitative researcher was used for cognitive interviews and expert input sought (e.g., seven members of the multidisciplinary pediatric burn team at a Canadian site) to revise scales after each round of cognitive interviews. Additionally, the authors transparently and data-driven manner present key domains and participant quotes to support inclusion of the SCAR-Q domains (i.e., scar appearance, scar symptoms, and psychosocial impact) and items.

Moreover, upon further review of the field test study, you determine that the authors demonstrate the hypothesis-based construct validity of the SCAR-Q through creation of clinically- and evidence-driven hypotheses and demonstrating a correlation with another scar-related PROM called the Patient and Observer Scar Assessment Scale (POSAS).

Is the PROM Reliable?

You determine that the field test of the SCAR-Q confirmed the psychometric properties of this PROM through Rasch measurement theory (RMT) analysis [65]. Specifically, for all domains of the SCAR-Q—scar appearance, scar symptoms, and psychosocial impact—adequate internal consistency reliability (Cronbach alpha values of 0.96, 0.91, and 0.95, respectively) and test-retest reliability (intraclass correlation coefficient values of 0.92, 0.94, and 0.88, respectively) were supported [65].

Is the PROM Responsive to Change?

While the development and field test study establish the psychometric properties of the SCAR-Q, they do not examine responsiveness to change. You review a follow-up study by Ziolkowski et al. which demonstrates that SCAR-Q scores change with the age of the scar [66]. This is reassuring; however, these studies are limited by their cross-sectional design.

Ultimately, you conclude that, compared to other commonly used PROMs assessing surgical and traumatic scars [1]—such as the Bock Quality of Life Questionnaire for Patients with Keloid and Hypertrophic Scarring (Bock QoL), the Patient and Observer Scar Assessment Scale (POSAS; including the mPOSAS version), the Patient-Reported Impact of Scars Measure (PRISM), and the Patient Scar Assessment Questionnaire (PSAQ)—the SCAR-Q demonstrates robust evidence of patient-driven development, validity, and reliability in patients with surgical scars. You conclude that the SCAR-Q is directly applicable to your proposed study population and recommend it as the primary outcome measure for your proposed RCT. Notably, a current limitation of the SCAR-Q is that there has yet to be a published minimal important difference that researcher may use to interpret the results.

Readers are also encouraged to refer to Chap. 10

Resolution of the Clinical Scenario

In checking the COSMIN Database of Systematic Reviews, one can readily find the SCAR-Q. Based on your critical appraisal, you believe the SCAR-Q meets the criteria of a rigorously developed PROM. The SCAR-Q is suitable to measure scar appearance, scar symptoms, and psychosocial impact in the burn population and you recommend that it be included in your proposed RCT.

Summary

Developing a patient-reported outcome measure (PROM) can be a time- and resource-intensive task that involves multiple decision points along the way. A well-developed PROM, with demonstrated content validity, reliability, validity, and responsiveness in the target population, has the power to meaningfully incorporate the patient voice into treatment

decision-making and outcome assessment. Ultimately, the uptake of PROMs contributes to making health care patient-centered, high quality, and high value.

Appendix

Search Strategy on PubMed Medline

Keywords: “Burn” AND “Scar” AND “Patient reported outcome measure”

Filter: ‘Systematic reviews’; last 10 years (January 2014—November 2024).

Search Results

1. Mundy LR, Miller HC, Klassen AF, Cano SJ, Pusic AL. Patient-reported outcome instruments for surgical and traumatic scars: a systematic review of their development, content, and psychometric validation. *Aesthetic Plast Surg.* 2016;40:792–800.
2. Carrière ME, Kwa KA, de Haas LE, Pijpe A, Tyack Z, Ket JC, van Zuijlen PPM, de Vet HCW, Mokkink LB. Systematic review on the content of outcome measurement instruments on scar quality. *Plast Reconstr Surg Glob Open.* 2019;7(9):e2424.
3. Harris IM, Lee KC, Deeks JJ, Moore DJ, Moiemien NS, Dretzke J. Pressure-garment therapy for preventing hypertrophic scarring after burn injury. *Cochrane Database Syst Rev.* 2024(1).
4. Vosinakis C, Ippoliti S, Samoladas E, Haidich A-B, Gamatsi IE, Smith L, Pourzitaki C. Effectiveness of hand reconstruction techniques for the treatment of postburn contractures of the hand: a systematic review. *Burns.* 2024.
5. Mahar PD, Spinks AB, Cleland H, Bekhor P, Waibel JS, Lo C, Goodman G. Improvement of burn scars treated with fractional ablative CO2 lasers—a systematic review and meta-analysis using the Vancouver Scar Scale. *J Burn Care Res.* 2021;42(2):200–6.

References

- Mundy LR, Miller HC, Klassen AF, Cano SJ, Pusic AL. Patient-reported outcome instruments for surgical and traumatic scars: a systematic review of their development, content, and psychometric validation. *Aesth Plast Surg*. 2016;40:792–800.
- Carrière ME, Kwa KA, de Haas LE, Pijpe A, Tyack Z, Ket JC, van Zuijlen PPM, de Vet HCW, Mekkink LB. Systematic review on the content of outcome measurement instruments on scar quality. *Plast Reconstr Surg Glob Open*. 2019;7(9):e2424.
- Harris IM, Lee KC, Deeks JJ, Moore DJ, Moiemien NS, Dretzke J. Pressure-garment therapy for preventing hypertrophic scarring after burn injury. *Cochrane Database Syst Rev*. 2024;1
- Vosinakis C, Ippoliti S, Samoladas E, Haidich A-B, Gamatsi IE, Smith L, Pourzitaki C. Effectiveness of hand reconstruction techniques for the treatment of postburn contractures of the hand: a systematic review. *Burns*. 2024;50:107281.
- Mahar PD, Spinks AB, Cleland H, Bekhor P, Waibel JS, Lo C, Goodman G. Improvement of burn scars treated with fractional ablative CO2 lasers—a systematic review and meta-analysis using the Vancouver Scar Scale. *J Burn Care Res*. 2021;42(2):200–6.
- Klassen AF, Ziolkowski N, Mundy LR, Miller HC, McIlvride A, DiLaura A, Fish J, Pusic AL. Development of a new patient-reported outcome instrument to evaluate treatments for scars: the SCAR-Q. *Plast Reconstr Surg Glob Open*. 2018;6(4):e1672.
- Black N, Burke L, Forrest CB, Ravens Sieberer U, Ahmed S, Valderas J, Bartlett SJ, Alonso J. Patient-reported outcomes: pathways to better health, better services, and better societies. *Qual Life Res*. 2016;25:1103–12.
- Lavallee DC, Chenok KE, Love RM, Petersen C, Holve E, Segal CD, Franklin PD. Incorporating patient-reported outcomes into health care to engage patients and enhance care. *Health Aff*. 2016;35(4):575–82.
- Makhni EC. Meaningful clinical applications of patient-reported outcome measures in orthopaedics. *JBJS*. 2021;103(1):84–91.
- Rikkert MGO, van der Wees PJ, Schoon Y, Westert GP. Using patient reported outcomes measures to promote integrated care. *Int J Integr Care*. 2018;18(2)
- Hughes SE, Aiyegbusi OL, McMullan C, Turner GM, Anderson N, Cruz Rivera S, Collis P, Glasby J, Lasserson D, Calvert M. Patient-reported outcomes in integrated health and social care. *JRSM Open*. 2024;15(3)
- Administration USDoHaHSFaD. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims; 2009. Available from: <https://www.fda.gov/media/77832/download>.
- De Vet HC, Terwee CB, Mekkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge University Press; 2011.
- Jensen MP, Karoly P, Braver S. The measurement of clinical pain intensity: a comparison of six methods. *Pain*. 1986;27(1):117–26.
- Ahmad F, Jhaji AK, Stewart DE, Burghardt M, Bierman AS. Single item measures of self-rated mental health: a scoping review. *BMC Health Serv Res*. 2014;14:1–11.
- Smets E, Garssen B, Bonke B, De Haes J. The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. *J Psychosom Res*. 1995;39(3):315–25.
- Hays RD, Reise S, Calderón JL. How much is lost in using single items? *J Gen Intern Med*. 2012;27:1402–3.
- McKenna SP. Measuring patient-reported outcomes: moving beyond misplaced common sense to hard science. *BMC Med*. 2011;9:1–12.
- Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, Ring L. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1—eliciting concepts for a new PRO instrument. *Value Health*. 2011;14(8):967–77.
- Brown J. Interviews, focus groups, and Delphi techniques, *Advanced research methods for applied psychology*. Routledge; 2018. p. 95–106.
- Roulston K, Choi M. Qualitative interviews. In: *The SAGE handbook of qualitative data collection*; 2018. p. 233–49.
- Siedlecki SL. Conducting interviews for qualitative research studies. *Clin Nurse Spec*. 2022;36(2):78–80.
- Dilley P. Conducting successful interviews: tips for intrepid research. *Theory Pract*. 2000;39(3):131–7.
- Edwards R, Holland J. What is qualitative interviewing? Bloomsbury Academic; 2013.
- Brinkmann S. Qualitative interviewing. Oxford University Press; 2013.
- Turner DW III. Qualitative interview design: a practical guide for novice investigators. *Qual Rep*. 2010;15(3):754.
- Klassen AF, Kaur M, Johnson N, Kreukels BP, McEvenue G, Morrison SD, Mullender MG, Poulsen L, Ozer M, Rowe W, Satterwhite T, Savard K, Semple J, Sørensen JA, van de Grift TC, van der Meij-Ross M, Young-Afat D, Pusic AL. International phase I study protocol to develop a patient-reported outcome measure for adolescents and adults receiving gender-affirming treatments (the GENDER-Q). *BMJ Open*. 2018;8(10):e025435.
- Picchietti DL, Abuckle RA, Abetz L, Durmer JS, Ivanenko A, Owens JA, Croenlein J, Allen RP, Walters AS. Pediatric restless legs syndrome: analysis of symptom descriptions and drawings. *J Child Neurol*. 2011;26(11):1365–76.
- Oluboyede Y, Hulme C, Hill A. Development and refinement of the WAIE: a new obesity-specific quality of life measure for adolescents. *Qual Life Res*. 2017;26:2025–39.

30. Mpundu-Kaambwa C, Bulamu NB, Lines L, Chen G, Whitehurst DG. Exploring the use of pictorial approaches in the development of paediatric patient-reported outcome instruments: a systematic review. *PharmacoEconomics*. 2024;42(1):S9–S37.
31. Hu S, Liu J, Cornacchi SD, Klassen AF, Pusic AL, Kaur MN. Extracting big data from the internet to support the development of a new patient-reported outcome measure for breast implant illness: a proof of concept study. *Qual Life Res*. 2024;33(7):1975–83.
32. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. *Nurs Health Sci*. 2013;15(3):398–405.
33. Humble N, Mozelius P. Content analysis or thematic analysis: similarities, differences and applications in qualitative research. In: European conference on research methodology for business and management studies; 2022.
34. Neuendorf KA. Content analysis and thematic analysis, Advanced research methods for applied psychology. Routledge; 2018. p. 211–23.
35. Elo S, Kyngäs H. The qualitative content analysis process. *J Adv Nurs*. 2008;62(1):107–15.
36. Terry G, Hayfield N, Clarke V, Braun V. Thematic analysis. In: *The SAGE handbook of qualitative research in psychology*, vol. 2(17–37); 2017. p. 25.
37. Administration USDoHaHSFaD. Patient-focused drug development: methods to identify what is important to patients guidance for industry, Food and Drug Administration staff, and other stakeholders; 2022. Available from: <https://www.fda.gov/media/131230/download>.
38. Khadka J, Gothwal VK, McAlinden C, Lamoureux EL, Pesudovs K. The importance of rating scales in measuring patient-reported outcomes. *Health Qual Life Outcomes*. 2012;10:1–13.
39. Terwee CB, Prinsen C, Chiarotto A, De Vet H, Bouter LM, Alonso J, Westerman MJ, Patrick DL, Mokkink LB. COSMIN methodology for assessing the content validity of PROMs—user manual, vol. 27. Amsterdam: VU University Medical Center; 2018. p. 1159–70.
40. Willis GB. Cognitive interviewing: a tool for improving questionnaire design. Sage Publications; 2004.
41. Christodoulou C, Junghaenel DU, DeWalt DA, Rothrock N, Stone AA. Cognitive interviewing in the evaluation of fatigue items: results from the patient-reported outcomes measurement information system (PROMIS). *Qual Life Res*. 2008;17(10):1239–46.
42. Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, Ring L. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2—assessing respondent understanding. *Value Health*. 2011;14(8):978–88.
43. Linacre JM. Sample size and item calibration stability. *Rasch Meas Trans*. 1994;7:328.
44. Birnbaum A, Lord F, Norvick M. Statistical theory of mental test scores. Preprint at 1968.
45. Wu M, Tam HP, Jen T-H, Wu M, Tam HP, Jen T-H. Classical test theory. In: *Educational measurement for applied researchers: theory into practice*; 2016. p. 73–90.
46. Traub RE. Classical test theory in historical perspective. *Educ Meas*. 1997;16:8–13.
47. DeVellis RF. Classical test theory. *Med Care*. 2006;44(11):S50–S9.
48. Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clin Ther*. 2014;36(5):648–62.
49. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess*. 2009;13.
50. Andrich D, Marais I. A course in Rasch measurement theory. In: *Measuring in the educational, social and health sciences*; 2019. p. 41.
51. Van der Linden WJ. Handbook of item response theory: three volume set. CRC Press; 2018.
52. Cella D, Chang C-H, Heinemann AW. Item response theory (IRT): applications in quality of life measurement, analysis and interpretation. In: *Statistical methods for quality of life studies: design, measurements and analysis*. Springer; 2002. p. 169–85.
53. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care*. 2000;38(9):II-28–42.
54. Petrillo J, Cano SJ, McLeod LD, Coon CD. Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value Health*. 2015;18(1):25–34.
55. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res*. 2010;19:539–49.
56. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Qual Life Res*. 2021;30:2197–218.
57. McHorney CA. Measuring and monitoring general health status in elderly persons: practical and methodological issues in using the SF-36 Health Survey. *The Gerontologist*. 1996;36(5):571–83.
58. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008;61(2):102–9.
59. Wang Y, Devji T, Qasim A, Hao Q, Wong V, Bhatt M, Prasad M, Wang Y, Noori A, Xiao Y, Ghadimi M, Lozano LEC, Phillips MR, Carrasco-Labra A, King M, Terluin B, Terwee CB, Walsh M, Furukawa TA, Guyatt GH. A systematic survey identified method-

- ological issues in studies estimating anchor-based minimal important differences in patient-reported outcomes. *J Clin Epidemiol*. 2022;142:144–51.
60. Wang Y, Devji T, Carrasco-Labra A, King MT, Terluin B, Terwee CB, Walsh M, Furukawa TA, Guyatt GH. A step-by-step approach for selecting an optimal minimal important difference. *BMJ*. 2023;381.
61. INstruments CC-bSftsohM. Available from: <https://www.cosmin.nl>.
62. Prinsen CA, Mokkink LB, Bouter LM, Alonso J, Patrick DL, De Vet HC, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27:1147–57.
63. Dobbs TD, Hughes S, Mowbray N, Hutchings HA, Whitaker IS. How to decide which patient-reported outcome measure to use? A practical guide for plastic surgeons. *J Plast Reconstr Aesthet Surg*. 2018;71(7):957–66.
64. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10(4):407–15.
65. Ziolkowski NI, Pusic AL, Fish JS, Mundy LR, She RW, Forrest CR, Hollenbeck S, Arriagada C, Calcagno M, Greenhalgh D, Klassen AF. Psychometric findings for the SCAR-Q patient-reported outcome measure based on 731 children and adults with surgical, traumatic, and burn scars from four countries. *Plast Reconstr Surg*. 2020;146(3):331e–8e.
66. Ziolkowski NI, Behman R, Klassen AF, Fish JS, Mundy LR, She RW, Forrest CR, Hollenbeck S, Arriagada C, Greenhalgh D, Pusic AL. Determining the independent risk factors for worse SCAR-Q scores and future SCAR revision surgery. *Plast Reconstr Surg*. 2021;148(1):203–12.

Surrogate Endpoints

9

Dan Cohen, Japleen Thind, Nicole Simunovic,
and Olufemi Ayeni

Abbreviations

AIDS	Acquired immunodeficiency syndrome
BMD	Bone mineral density
CVD	Cardiovascular disease
CGD	Chronic granulomatous disease
ERT	Estrogen replacement therapy
HIV	Human immunodeficiency virus
O-2	Superoxide production
RCT	Randomized controlled trial
STE	Surrogate threshold effect
T4	Thyroxine
TH	Total hip
T3	Triiodothyronine

Clinical Scenario

You are the orthopedic surgery resident on-call at your hospital and are consulted by the emergency department to see an 82-year-old lady who sustained an acute left intertrochanteric hip fracture after a mechanical fall. The patient has a known

past medical history significant for hypertension, dyslipidemia, and gastroesophageal reflux disease. On further questioning, the patient notes that she sustained a previous distal radius fracture on the right side approximately 2 years ago which thankfully healed without the need for any surgical intervention.

The patient is then consented to surgery and undergoes open reduction and internal fixation of her left hip in a timely manner. The next day during your morning rounds, you go to see the patient and she asks you: “Doctor, I never broke a bone in my entire life, and this is the second bone I broke in the last 2 years. Why is this happening and is there anything I can do to prevent this from happening again? My family friend who is a doctor in the United States told me I should start taking a bisphosphonate.” You know that bisphosphonates do have the ability to preserve bony density by preventing bone resorption from occurring; however, you are not sure if improved bone mineral density has any association with a reduction in fracture risk. As such, you tell the patient that you will get back to them at morning rounds tomorrow after you have searched the literature in order to provide them with an evidence-based recommendation.

D. Cohen · N. Simunovic · O. Ayeni (✉)
Division of Orthopedic Surgery, Department of
Surgery, McMaster University,
Hamilton, ON, Canada
e-mail: dan.cohen@medportal.ca;
simunon@mcmaster.ca; ayenif@mcmaster.ca

J. Thind
Faculty of Health Sciences, McMaster University,
Hamilton, ON, Canada
e-mail: japleen.thind@mail.utoronto.ca

Introduction

Endpoints are measurable clinical and biological findings that are used for the development and assessment of therapeutic interventions. A clinical

endpoint is a direct measure of clinical benefit from an intervention, including recovery, survival, or reduced adverse events [1]. A clinical endpoint is considered the most credible indicator of treatment effect and target outcomes of ultimate interest that reflect how a patient feels, functions, or survives [1]. Surrogate endpoints substitute for the treatment effects of an intervention on the target outcomes. Some surrogate endpoints are biomarkers that indicate biological processes or responses to therapeutic interventions [2]. The concept of surrogacy in clinical trials can be broader and include intermediate outcomes that measure function or symptoms, such as exercise tolerance or angina frequency, and are used as substitute for target outcomes [3, 4]. Generally, surrogate endpoints are used in trials as a substitute measure for how a patient may feel, function, or how long they survive. As such, surrogate endpoints do not directly measure the primary clinical benefit of interest but rather predicts the clinical benefit or harm based on scientific evidence [4].

Advantages of Surrogate Endpoints

The use of surrogate endpoints is becoming increasingly common in clinical research due to their cost-effectiveness and ease of measure compared to primary endpoints. For instance, mammographic density can be detected using X-rays and is positively associated with an increased risk of breast cancer [5]. Mammographies are typically more cost-effective in screening for breast cancer compared to a diagnostic test such as a biopsy. An advantage of some surrogate endpoints is that they can be quantified and help in clinical decision-making. In the mid-1990s, a new methodology was developed to quantify HIV RNA, also known as viral load, in patient serum. Validation studies have proven viral load to be a meaningful surrogate endpoint for HIV disease progression and death. In patients with HIV with an uncontrolled viral load and low CD4 count, there is an increased risk of postoperative mortality and complications. Hence, for elective surgeries, it is recommended that patients have a viral load level <200 copies/mL and a CD4 count

>200 cells/mm³. Since viral load and CD4 count are objective surrogate endpoints for HIV disease progression, they help surgeons consult patients with HIV [6].

In general, in patients with HIV, the combination of a low CD4 count and an uncontrolled viral load is associated with an increased risk for postoperative mortality and complications. If surgery is elective and the patient has a viral load level ≥ 200 copies/mL or CD4 count ≤ 200 cells/mm³, clinicians should consult with the patient's primary care provider or an experienced HIV care provider.

Furthermore, surrogate endpoints reduce the sample size required to observe an effect. For instance, in studies on cardiovascular disease (CVD), a common patient-important primary endpoint is mortality. A study analyzed CVD outcomes and mortality among 24,014 men and women aged 39–79 years. The overall ratio of 10-year CVD to 10-year CVD mortality was 5.4. Further, stratification by age and gender found ratios of 11.7 and 28.5 in young (39–50 years) men and women, and ratios of 2.4 and 3.2 in older (75–79 years) men and women, respectively [7]. These ratios suggest a higher likelihood of survival after experiencing a CVD event than death. From a research perspective, a definitive trial with death as the final outcome would involve thousands of patients followed up for several years. Alternatively, a trial with surrogate endpoints, such as carotid artery intima-media thickness or luminal loss, would require a smaller sample size and a shorter follow-up period [8].

Sometimes, there are ethical concerns with measuring clinical endpoints. In clinical practice, plasma acetaminophen concentration is used as a pharmacological biomarker to evaluate if treatment is necessary for acetaminophen overdose. Alternatively, it would be unethical and harmful to the patient to wait until evidence of liver damage before prescribing treatment [9]. In the context of research, measuring clinical endpoints such as overall survival can prolong a study, therefore delaying patient access to potentially beneficial treatments. A study showed that using surrogate endpoints in oncology can shorten the trial duration by 11–19 months, further reducing the cost of a clinical trial [10].

Limitations of Surrogate Endpoints

Smaller sample sizes and shorter durations of clinical trials are associated with greater uncertainty regarding the effectiveness of the intervention. Figure 9.1 illustrates conceptual models and examples of limitations of surrogate endpoints. Despite a strong correlation between the surrogate endpoint and disease progression, it may not lie in the same pathophysiological process that results in the clinical endpoint (Fig. 9.1a). Estrogen replacement therapy (ERT) is often prescribed for the treatment of menopausal symptoms and the prevention of osteoporosis [14]. Menopausal women are at high risk of CVD. It is also well established that ERT favor-

ably affects serum lipid profile, which affects the genesis and progression of CVD [15]. However, ERT does not reduce CVD, and in some studies has rather been associated with increased risk [15–17]. Thus, ERT and serum lipid profile are associated with CVD but lie in different pathophysiological pathways.

The intervention may only affect the pathway mediated through the surrogate endpoint, resulting in false positives (Fig. 9.1b). Flosequinan was used to treat congestive heart failure due to improvement in exercise tolerance associated with its use. The subsequent PROFILE study provided significant evidence that flosequinan increased total mortality [11]. A false negative may happen when the intervention only affects

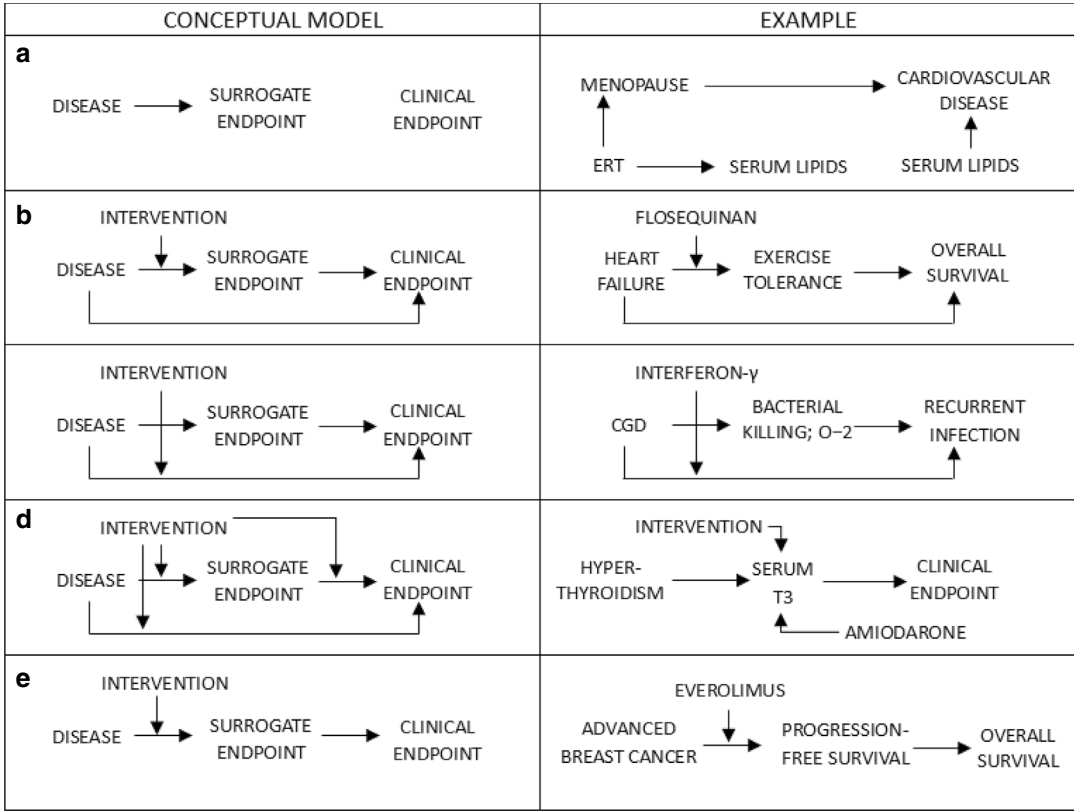


Fig. 9.1 Limitations of surrogate endpoints. (a) Surrogate endpoint does not lie in the same pathophysiological pathway. (b) Intervention only affects the surrogate endpoint. (c) Intervention only affects the clinical endpoint. (d) Intervention interacts with the pathophysiological pathway in unintended and unrecognized

mechanisms. (e) Incorrect estimation of the intervention’s magnitude of effect on the clinical endpoint. *Abbreviations:* ERT estrogen replacement therapy, CGD chronic granulomatous disease, O-2 superoxide production. (Adapted from Packer et al. [11], Fleming et al. [12] and Weintraub et al. [13])

pathways independent of the surrogate endpoint (Fig. 9.1c). The International CGD Study Group designed a trial to investigate the effects of interferon- γ in preventing recurrent infections in chronic granulomatous disease. Before initiation, the study was expanded to investigate the long-term effects of interferon- γ on the clinical endpoint of recurrent infections. During the trial, interferon- γ showed no detectable effect on the biomarkers (bacterial killing and superoxide production), but it was associated with a statistically and clinically significant reduction in the rate of recurrent infections [12, 18]. A shorter trial investigating the effects of interferon- γ only on superoxide production and bacterial killing would have resulted in misleading conclusions regarding an effective treatment.

It is also possible for the intervention to affect the final outcomes through unexpected or unidentified mechanisms of action that are independent of the disease pathway (Fig. 9.1d). This is when different variables in a clinical trial are affected by confounders. Serum T3 is a marker of the tissue damage that thyroid hormone causes in patients with hyperthyroidism. However, its usefulness is blunted in patients taking amiodarone, which interferes with the conversion of T4 to T3 without necessarily altering thyroid function. Hence, amiodarone would be a major confounder in a clinical trial that aims to use serum T3 as a surrogate endpoint in patients with hyperthyroidism [2].

A surrogate endpoint may lie in the same pathophysiological pathway as the disease and may capture entirely the intervention's effect on the disease, but the magnitude of the effect may be over- or underestimated (Fig. 9.1e). Trials using surrogate endpoints have been shown to overestimate the health benefits of interventions by >40%, compared to trials using clinical endpoints [19]. Overestimation results from the reduced power of studies with smaller sample sizes. The BOLERO-2 study evaluated the efficacy of everolimus in hormone receptor-positive advanced breast cancer. The primary objective of the study was the surrogate endpoint of progression-free survival, everolimus plus exemestane more than doubled progression-free

survival versus placebo plus exemestane in postmenopausal women with advanced breast cancer [20]. Continuation of the study to evaluate the secondary objective of overall survival showed no statistical improvement in survival from everolimus, thus the initial trial overestimated the efficacy of everolimus [21].

Validation of Surrogate Endpoints

Given the limitations of surrogate endpoints and their consequences in administering ineffective treatments, the relationship between surrogate endpoints and clinical endpoints must be validated appropriately. Many statistical strategies have been proposed to validate surrogate endpoints, but their complexity affects their use in research. Ciani et al. proposed a three-step methodological framework. The first step is to establish the level of evidence available. According to the framework, level 3 evidence for a surrogate endpoint is biological plausibility and level 2 evidence is when a strong association exists between the surrogate and the final endpoint across cohorts or at the level of the individual patient. Level 1 evidence is established in randomized controlled trials that show that intervention-related changes in the surrogate are associated with intervention-related changes in the clinical endpoint [22, 23]. The second step is to assess the strength of the association between the surrogate and clinical endpoints using statistical approaches. A meta-analysis of all randomized controlled trials of an intervention is performed for level 1 surrogacy. The final step is to quantify the expected effect on the clinical endpoint based on the observed effect on the surrogate endpoint. This quantification can be done using regression equations for the trial-level surrogacy, Bayesian multivariate meta-analyses, or surrogate threshold effect (STE) [24]. As discussed earlier, poorly validated surrogate outcomes can be misleading with regard to their clinical or economic benefits, posing the risk of potentially harmful interventions entering the market. Validation of surrogate endpoints to establish true surrogacy overcomes these limitations and produces meaningful outcomes [25].

Literature Search: Bone Mineral Density (BMD) and Fracture Outcomes

To resolve the clinical scenario, you formulate the following research question: Are bisphosphonates associated with fracture risk reduction? You conduct your literature search using the search engine PubMed and the search strategy described in Appendix 1. The search strategy yielded 54 results, of which only 2 validation studies were meta-analyses that established trial-level or level 1 surrogacy between BMD and fracture risk reduction with use of bisphosphonates. The chosen study quantified the association between BMD and fracture outcomes using STE, while the other validation study calculated the trial-level correlation coefficient. You recognize the limitations of trial-level correlation coefficient [26], and choose to address the clinical scenario using the study by Eastell et al. titled “Validation of the Surrogate Threshold Effect for Change in Bone Mineral Density as a Surrogate Endpoint for Fracture Outcomes: The FNIH-ASBMR SABRE Project” [27].

What Are the Results?

Eastell et al. included 16 trials, most of which predominantly enrolled women. The bisphosphonate drugs tested in nine trials were alendronate, ibandronate oral, risedronate, and ibandronate iv. Some studies focused on other drugs used for fracture risk reduction, such as denosumab and lasofoxifene, but the most relevant to this clinical scenario are the bisphosphonates. Four STE estimates were calculated for four fracture categories: all fractures, vertebral fracture, hip fracture, and nonvertebral fracture. You bring your attention to the significantly powered trials that were associated with a change in total hip BMD at 24 months exceeding the STE, and a significant reduction in fracture risk ($p < 0.05$) was noted. Across 61,415 participants followed up for 24 months, a mean net difference in BMD ranging from 1.3% to 5.4% was recorded with the use of bisphosphonates, with progressively higher fracture risk reduction with a higher STE.

Are the Results Valid?

To confirm the validity of BMD as a surrogate endpoint for fracture outcomes, you use the three-step methodological framework established by Ciani et al. Firstly, you confirm that level 1 evidence or trial level surrogacy has been established for BMD and fracture outcomes. In the chosen study, nine RCTs show that treatment-related changes in the surrogate (i.e. BMD) are associated with treatment-related changes in the clinical endpoint (i.e. fracture outcomes). Secondly, Eastell et al. conducted a meta regression to find the strength of association between treatment-related changes in BMD and fracture risk reduction, which was quantified by calculating the STE. Specifically, the STE is defined as the minimum change in the level of a certain surrogate marker that would predict improvement in a particular outcome measure with 95% certainty. Using the STE you can be sure that the results are valid and that the bisphosphonates that exceed the STE for fracture risks will be effective.

Are the Results Applicable to My Patients?

The chief concern of the patient is to reduce their risk of fractures. From the chosen study, we know that preserving BMD with bisphosphonates is associated with a significant reduction in fracture risk. It is also specified in the article that most trials predominantly included women except for two trials. Among the excluded studies were studies on participants with certain conditions (e.g., glucocorticoid-induced osteoporosis), and participants consuming strontium ranelate. The age of participants across the trials along with other demographic information was not included. The authors standardized the definition of fractures and included patients who sustained fragility fractures which are generally defined as lower energy fracture mechanisms that would otherwise have not resulted in a fracture in a young

and healthy patient. You are confident that the results of this study address the clinical scenario and some aspects of the study are comparable with your patient's experience (i.e., the patient is a female experiencing fragility fractures). However, you know that patient's past medical history of hypertension, dyslipidemia, and gastroesophageal reflux disease can potentially affect the effectiveness of the bisphosphonates. Since the patient has had surgical intervention, which is associated with great risk for someone her age, you decide to proceed with the bisphosphonates with caution to prevent more fractures.

Resolution of Clinical Scenario

You return to your patient and explain to them that there is a significant reduction in fragility fracture risk in patients who take bone-preserving medications such as bisphosphonates. You explain to the patient that utilizing the STE, a change in BMD ranging from 3.2% to 5.8% results in a progressive reduction in the incidence of hip fractures. Before giving the patient a prescription, you inform her of the potential side effects. Firstly, oral bisphosphonates are correlated with upper gastrointestinal adverse effects, including gastrointestinal reflux, gastritis, esophageal or gastric ulcers, and esophagitis. Since the risk of upper gastrointestinal adverse effects is slightly lower with risedronate compared to alendronate, you decide to prescribe oral risedronate [28]. You advise her to stay upright for at least 30 min after taking the drug on an empty stomach or until after first food intake to reduce esophageal discomfort and ulceration [28]. Given this patient's past medical history, you ask her to seek medical attention if she experiences any esophageal discomfort. You also recommend that she takes supplemental calcium and vitamin D to ensure adequate absorption of the drug [28]. Before initiating risedronate, you check her serum calcium, vitamin D, and renal function. You plan on evaluating her BMD levels in 1–2 years after initiating risedronate.

Conclusion

Surrogate endpoints are used extensively in clinical research due to their efficiency and cost effectiveness. However, their limitations must be considered before making clinical decisions based on trials involving surrogate endpoints. Using unvalidated surrogate endpoints in research or clinical practice poses a public health risk. Hence, it is important to understand how to identify, critically appraise, and apply validated surrogate outcome studies in practice.

Appendix 1

The keywords “bone mineral density, BMD, bisphosphonates, fracture outcome, fracture risk” were entered into the search engine. The first two and last keywords were separated by the OR Boolean operator and the rest were combined with the AND Boolean operator. To ensure that studies were limited to validation studies of BMD as a surrogate endpoint, the keyword “valid” was included in the search using the Boolean operator AND. The results were limited to studies published in the past 10 years (2013–2023).

References

1. Roever L. Endpoints in clinical trials: advantages and limitations. *Evid Based Med Pract*. 2015;1(2):1000e111. <https://doi.org/10.4172/ebmp.1000e111>.
2. Aronson JK. Biomarkers and surrogate endpoints. *Br J Clin Pharmacol*. 2005;59(5):491–4. <https://doi.org/10.1111/j.1365-2125.2005.02435.x>.
3. Korn EL, Freidlin B. Surrogate and intermediate endpoints in randomized trials: what's the goal? *Clin Cancer Res*. 2018;24(10):2239–40. <https://doi.org/10.1158/1078-0432.CCR-18-0183>.
4. Ciani O, Manyara AM, Davies P, Stewart D, Weir CJ, Young AE, Blazeby J, Butcher NJ, Bujkiewicz S, Chan A, Dawoud D, Offringa M, Ouwers M, Hróbjartsson A, Amstutz A, Bertolaccini L, Bruno VD, Devane D, Faria CDCM, Gilbert BP, Harris R, Lassere M, Marinelli L, Markham S, Powers JH, Rezaei Y, Richert L, Schwendicke F, Tereshchenko LG, Thoma A, Turan A, Worrall A, Christensen R, Collins GS, Ross JS, Taylor RS. A framework for the definition and interpretation of the use of surrogate

- endpoints in interventional trials. *eClinicalMedicine*. 2023;65:102283. <https://doi.org/10.1016/j.eclinm.2023.102283>.
5. Fabian C. Surrogates are just surrogates, but helpful just the same. *Breast Cancer Res*. 2007;9(2):S18. <https://doi.org/10.1186/bcr1816>.
 6. Yan L, Ellman T, McNairy M, McGowan JP, Fine SM, Vail RM, Merrick ST, Radix AE, Hoffmann CJ, Gonzalez CJ. Perioperative care in adults with HIV. New York State Department of Health AIDS Institute (NYSDOH AI); 2021. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK576018/>. Accessed 19th January 2024
 7. Jørstad HT, Colkesen EB, Boekholdt SM, Tijssen JG, Wareham NJ, Khaw KT, Peters RJ. Estimated 10-year cardiovascular mortality seriously underestimates overall cardiovascular risk. *Heart*. 2016;102(1):63–8. <https://doi.org/10.1136/heartjnl-2015-307668>.
 8. Wittes J, Lakatos E, Probstfield J. Surrogate endpoints in clinical trials: cardiovascular diseases. *Stat Med*. 1989;8(4):415–25. <https://doi.org/10.1002/sim.4780080405>.
 9. Vliegthart ADB, Antoine DJ, Dear JW. Target biomarker profile for the clinical management of paracetamol overdose. *Br J Clin Pharmacol*. 2015;80(3):351–62. <https://doi.org/10.1111/bcp.12699>.
 10. Chen EY, Joshi SK, Tran A, Prasad V. Estimation of study time reduction using surrogate end points rather than overall survival in oncology clinical trials. *JAMA Intern Med*. 2019;179(5):642–7. <https://doi.org/10.1001/jamainternmed.2018.8351>.
 11. Packer M, Pitt B, Rouleau JL, Swedberg K, DeMets DL, Fisher L. Long-term effects of flosequinan on the morbidity and mortality of patients with severe chronic heart failure: primary results of the PROFILE trial after 24 years. *JACC Heart Fail*. 2017;5(6):399–407. <https://doi.org/10.1016/j.jchf.2017.03.003>.
 12. Fleming TR. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med*. 1996;125(7):605. <https://doi.org/10.7326/0003-4819-125-7-199610010-00011>.
 13. Weintraub WS, Lüscher TF, Pocock S. The perils of surrogate endpoints. *Eur Heart J*. 2015;36:2212–8. <https://doi.org/10.1093/eurheartj/ehv164>.
 14. Boussein ML, Delmas PD. Considerations for development of surrogate endpoints for antifracture efficacy of new treatments in osteoporosis: a perspective. *J Bone Miner Res*. 2008;23(8):1155–67. <https://doi.org/10.1359/JBMR.080301>.
 15. Mosca L, Collins P, Herrington DM, Mendelsohn ME, Pasternak RC, Robertson RM, Schenck-Gustafsson K, Smith SC, Taubert KA, Wenger NK. Hormone replacement therapy and cardiovascular disease. *Circulation*. 2001;104(4):499–503. <https://doi.org/10.1161/hc2901.092200>.
 16. Sanchez RG, Sanchez Gomez LM, Carmona L, Figuls MR, i, Cosp XB. Hormone replacement therapy for preventing cardiovascular disease in post-menopausal women. *Cochrane Database Syst Rev*. 2005;(2):CD002229. <https://doi.org/10.1002/14651858.CD002229.pub2>.
 17. Bhargava S, de la Puente-Secades S, Schurgers L, Jankowski J. Lipids and lipoproteins in cardiovascular diseases: a classification. *Trends Endocrinol Metab*. 2022;33(6):409–23. <https://doi.org/10.1016/j.tem.2022.02.001>.
 18. The International Chronic Granulomatous Disease Cooperative Study Group*. A controlled trial of interferon gamma to prevent infection in chronic granulomatous disease. *N Engl J Med*. 1991;324(8):509–16. <https://doi.org/10.1056/NEJM199102213240801>.
 19. Ciani O, Manyara AM, Chan AW, Taylor RS, on behalf of the SPIRIT-SURROGATE/CONSORT-SURROGATE project group. Surrogate endpoints in trials: a call for better reporting. *Trials*. 2022;23(1):991. <https://doi.org/10.1186/s13063-022-06904-7>.
 20. Yardley DA, Noguchi S, Pritchard KI, Burris HA, Baselga J, Gnant M, Hortobagyi GN, Campone M, Pistilli B, Piccart M, Melichar B, Petrakova K, Arena FP, Erdkamp F, Harb WA, Feng W, Cahana A, Taran T, Lebwohl D, Rugo HS. Everolimus plus exemestane in postmenopausal patients with HR+ breast cancer: BOLERO-2 final progression-free survival analysis. *Adv Ther*. 2013;30(10):870–84. <https://doi.org/10.1007/s12325-013-0060-1>.
 21. Piccart M, Hortobagyi GN, Campone M, Pritchard KI, Lebrun F, Ito Y, Noguchi S, Perez A, Rugo HS, Deleu I, Burris HA 3rd, Provencher L, Neven P, Gnant M, Shivelband M, Wu C, Fan J, Feng W, Taran T, Baselga J. Everolimus plus exemestane for hormone-receptor-positive, human epidermal growth factor receptor-2-negative advanced breast cancer: overall survival results from BOLERO-2. *Ann Oncol*. 2014;25(12):2357–62. <https://doi.org/10.1093/annonc/mdu456>.
 22. Ciani O, Grigore B, Blommestein H, De Groot S, Möllenkamp M, Rabbe S, Daubner-Bendes R, Taylor RS. Validity of surrogate endpoints and their impact on coverage recommendations: a retrospective analysis across international health technology assessment agencies. *Med Decis Making*. 2021;41(4):439–52. <https://doi.org/10.1177/0272989X21994553>.
 23. Ciani O, Grigore B, Taylor RS. Development of a framework and decision tool for the evaluation of health technologies based on surrogate endpoint evidence. *Health Econ*. 2022;31(S1):44–72. <https://doi.org/10.1002/hec.4524>.
 24. Wheaton L, Papanikos A, Thomas A, Bujkiewicz S. Using Bayesian evidence synthesis methods to incorporate real world evidence in surrogate endpoint evaluation. *Med Decis Making*. 2023;43(5):539–52. <https://doi.org/10.1177/0272989X231162852>.
 25. Kemp R, Prasad V. Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? *BMC Med*. 2017;15:134. <https://doi.org/10.1186/s12916-017-0902-9>.

26. Baker SG. Five criteria for using a surrogate endpoint to predict treatment effect based on data from multiple previous trials. *Stat Med*. 2018;37(4):507–18. <https://doi.org/10.1002/sim.7561>.
27. Eastell R, Vittinghoff E, Lui LY, McCulloch CE, Pavo I, Chines A, Khosla S, Cauley JA, Mitlak B, Bauer DC, Bouxsein M, Black DM. Validation of the surrogate threshold effect for change in bone mineral density as a surrogate endpoint for fracture outcomes: the FNIH-ASBMR SABRE project. *J Bone Miner Res*. 2022;37(1):29–35. <https://doi.org/10.1002/jbmr.4433>.
28. Ganesan K, Goyal A, Roane D. Bisphosphonate. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2024. Available from <http://www.ncbi.nlm.nih.gov/books/NBK470248/>. Accessed 19th February 2024.



How to Assess an Article that Deals with Health-Related Quality of Life

10

Achilles Thoma, Jenny Santos,
Margherita Cadeddu, Eric Duku, Emily Dunn,
and Charles H. Goldsmith

Abbreviations

AFR	Autogenous Flap Reconstruction
CEA/CUA	Cost-Effectiveness Analysis/ Cost-Utility Analysis (CUA)
HRQL	Health-Related Quality of Life
IBR	Implant-Based Reconstructions
MCID/MID	Minimal Clinical Important Difference/Minimal Important Difference
PROM	Patient-Reported Outcome Measure
QALYs	Quality-Adjusted Life Years
QoL	Quality of Life
RCT	Randomized controlled trials (RCTs)

Introduction

Traditionally, the results of surgery have been determined by clinical outcomes such as prolonged survival and reduced morbidity. Death and survival rates are critical outcomes that are usually difficult to dispute, unless a patient is comatose in a vegetative state in an intensive care unit or on a neurosurgical ward. Surgeons and patients, however, can interpret morbidity outcomes differently. A hernia repair may be considered a success from the general surgeon's point of view but not so from the patient's point of view if the scar is painful due to skin scar neuromata or entrapment of deeper sensory nerves during the repair. Similarly, a

A. Thoma (✉)
Division of Plastic Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: athoma@mcmaster.ca

J. Santos
Southwestern Public Health (SWPH),
St. Thomas, ON, Canada

M. Cadeddu
Division of General Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada

E. Duku
Department of Health Research Methods,
Evidence, and Impact, McMaster University,
Hamilton, ON, Canada

Offord Centre for Child Studies, Department
of Psychiatry and Behavioral Neuroscience,
McMaster University, Hamilton, ON, Canada
e-mail: duku@mcmaster.ca

E. Dunn
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: dunne4@mcmaster.ca

C. H. Goldsmith
Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: goldsmith@mcmaster.ca

successful digit replant may be considered a success from a hand surgeon's point of view if the digit is viable but not from the patient's point of view if there is stiffness of the whole hand making it difficult to work, potentially leading to loss of employment or other negative outcomes.

Since the early 1990s, the outcomes research movement has shifted the declaration of a "successful surgical outcome" from surgeon to patient. In other words, the evaluation of outcomes in surgery have begun to take into account the patient's experiences, preferences, and values. This is, in essence, one of the tenets of what evidence-based surgery is all about (see Chap. 1). With the exception of some high-risk surgical subspecialties, such as cardiac surgery and oncological surgery, most of the surgical procedures we perform as surgeons are aimed at reducing morbidity or improving quality of life (QoL). The broader term QoL can be defined as "an individual's perception of their position in life in the context of the culture and values systems in which they live and in relation to their goals, expectations, standards, and concerns" [1]. A subcategory of QoL, health-related quality of life (HRQL), is defined by the Center of Disease Control and Prevention (CDC) as "an individual's or groups' perceived physical and mental health over time" [2]. Since around 1990, there has been an increasing trend to evaluate the outcomes of healthcare interventions using sophisticated HRQL scales (instruments, measures) [3, 4]. The same has been espoused for surgical interventions [5].

With time, we see more and more articles in surgical literature where the results of novel interventions are reported with HRQL scales or patient-reported outcome measures (PROMs) rather than the traditional physiologic outcomes. Decisions to adopt or reject novel interventions are based on such reports. It is therefore important for surgeons to be familiar with the appraisal of such articles that measure HRQL before they incorporate interventions into their practice and recommend them to their patients.

Clinical Scenario

A 47-year-old woman is seen at the multidisciplinary breast oncology clinic with a recent diagnosis of breast cancer found on mammogram and confirmed with biopsy. The cancer is localized to the breast and the recommendation is made for simple mastectomy. She would like to have post-mastectomy breast reconstruction. The plastic surgeon on the multidisciplinary team recommended implant-based reconstruction. The patient, however, has already explored the Internet and found out that there is a new method of breast reconstruction. It includes harvesting fat from hips and thighs via liposuction and making a new breast with it. She prefers this technique as one of her friends had breast reconstruction with this method and was happy with it. The plastic surgeon is not sure if the quality of life is better with this new technique but promises to review the literature and let her know on the next visit.

Literature Search

The ideal article addressing this surgical question would be a meta-analysis of randomized controlled trials (RCTs) comparing postmastectomy breast reconstruction with a two stage tissue expansion and autologous fat transfer.

In the absence of such a study design, one should look for a large single RCT that addresses the same question. Using the skills outlined by Banfield et al. in Chap. 4 and the Waltho et al. Users Guide to the Surgical Literature [6], one searches the literature after creating a clinical research question based on the PICOT format (see Chap. 3 on the clinical research question):

- *Population:* Patients who undergo mastectomy for breast cancer or prophylactic mastectomy
- *Intervention:* Breast reconstruction with autologous fat transfer
- *Comparative intervention:* Implant-based reconstruction
- *Outcome:* HRQL
- *Time horizon:* 12 months

From the components of the PICOT format above, the research question is: “In patients who undergo mastectomy for breast cancer, does post-mastectomy reconstruction with autologous fat transfer lead to better HRQL than implant-based reconstruction at 12 months”?

Using a filtered database, COCHRANE, we performed a search using the key words “breast reconstruction” AND “autologous fat transfer” AND “implant-based reconstruction” and “randomized controlled trial” and “quality of life” and limited the search to the last 10 years (2013–April 2024). The search returned three articles (see [Appendix](#)). One the studies was an RCT, that is, the BREAST trial, originally published by Piatowski et al. [7]. It appeared that the other two studies were based off that same RCT, as amendments, or sub-studies of the BREAST trial, one in which the primary outcome was breast sensibility or sensation and the other was cost-effectiveness. The original RCT, by Piatowski et al. [7], looked promising. Upon full-text review, we confirmed this was a study on postmastectomy breast reconstruction patients, in which participants were randomized to receive breast reconstruction with autologous fat transfer (AFT) or implant-based reconstruction (IBR). The primary outcome was QoL at 12 months after final surgery, as measured by the BREAST-Q questionnaire (see [Box 10.1](#)).

Box 10.1 Guidelines for the appraisal of an article in the surgical literature that purports to be reporting on health-related quality of life

A. Are the results valid?

Primary guides

- (i) Have the surgical investigators measured aspects of patients’ lives that surgical patients consider important?
- (ii) Have important aspects of HRQL been omitted?
- (iii) Are the HRQL instruments chosen valid, reliable, and responsive?

Secondary guides

- (i) Were HRQL assessments appropriately timed to evaluate the effects of the surgical intervention?
- (ii) If there were trade-offs between quantity and quality of life, did the investigators use an economic analysis?
- B. What were the results?**
 - (i) What was the magnitude of the effect on HRQL?
- C. Will the results help me in caring for my patients in my practice?**
 - (i) Will the information from this study help me inform my patients?

Primary Guides

Have the Investigators Measured Aspects of Patients’ Lives that Patients Consider Important?

Traditionally, surgeons decided if an intervention is successful based on the restoration of some anatomical defect, reverting laboratory data to normal range, successful resection of malignant disease based on imaging or some laboratory marker, or just simply by asking patients how they felt. The problem here is that we are not certain if we have considered all aspects of health that patients consider important. If one looks at the CDC definition of HRQL, one can see it is multidimensional [2]. One way to address this problem is to administer HRQL questionnaires to patients before and at various times after the surgical intervention and see if there has been a change.

The development of HRQL questionnaires (or PROMs) is an arduous process that could take years to develop. Such an instrument requires the combined effort of methodologists, psychologists, biostatisticians, patients, and other experts. According to the U.S. Food and Drug Administration (FDA), there are four measurable outcomes: patient-reported outcomes (PROs),

clinician-reported outcomes, observer-reported outcomes, and performance outcomes [8].

Streiner et al. [9] provide an excellent source on how to design health-measurement scales. As mentioned above, it takes a lot of effort to develop a new HRQL scale. Surgeons do not have to develop such scales unless one does not exist for a particularly common condition. McDowell provides many known QoL questionnaires and surgeons are encouraged to use this helpful resource [10]. The nomenclature of these HRQL instruments has evolved to patient-reported outcomes (PROs), a term that is used more frequently now [11]. Guidelines have been developed since 2010 for the development of a new patient-reported (PRO) instrument [12]. For PROMs or HRQL instruments to be valid, they need to have proper psychometric properties. By this, we mean that the PROMs reliably attach a numerical value to patients' feelings so they can be assessed. If they are not constructed by the correct methods they are invalid and unreliable. Some commonly used valid HRQL instruments can be found in Table 10.1. Although generic HRQL or PROMs provide some important information about both health status and health utility, they are not always the most appropriate choice, if chosen as the sole HRQL instrument [13]. If the quality of life is greatly affected by a specific disease, such

as in cancer patients, a disease-specific instrument should be used [13].

In appraising the Piatkowski et al. article, we need to assess whether the authors have measured aspects of patients' lives that the patients would consider important [7].

Indeed, patients who undergo mastectomy for breast cancer may experience negative effects on both physical and emotional well-being [14, 15]. They often opt to have breast reconstruction to alleviate pain, restore self-esteem, improve vitality, femininity, and sexuality [16–19]. Cumulative evidence over the last few years have shown that breast reconstruction has a positive effect on patients' quality of life [18, 20–22].

There are many techniques to reconstruct the mastectomy defect. These can be categorized into two main groups: implant-based reconstructions (IBR) and autogenous flap reconstruction (AFR). The AFR method of reconstruction, however, is more complex and often requires microsurgical techniques not available to many surgeons. The complications of these techniques are also variable [16, 23].

A procedure that transfers fat from an area of the body where it may be plentiful and grafts it to the chest wall to make a new permanent breast may be simpler and may have advantages over an implant-based reconstruction [24–26]. This

Table 10.1 Examples of commonly used disease-specific and generic HRQL instruments

Instrument	Abbreviation	Condition
Michigan Hand Questionnaire	MHQ	Any condition of the upper extremities
European Organization for Research and Treatment of Cancer Quality-of-Life Questionnaire Core 30	EORTC QLQ-C30	Generic cancer (more specific ones exist for different cancers)
Western Ontario and McMaster University osteoarthritis Index	WOMAC	Hip and knee osteoarthritis
Gastrointestinal Quality of Life Index	GIQLI	Gastrointestinal conditions
Short Form-36	SF-36	Generic: Health Status
EuroQol 5-Dimension	EQ-5D	Generic: Health Utility
Health Utility Index	HUI	Generic: Health Utility
Functional Assessment of Cancer Therapy Scale-Esophageal Cancer	FACT-E	Esophageal cancer
Functional Assessment of Cancer Therapy Scale-Head and Neck Cancer	FACT-HN	Head and neck cancer
Medical Outcomes Study Short-Form General Health Survey-20	MOS-SF20	Chronic conditions
Lymphoedema Quality of Life Tool	LYMQOL	Lymphoedema of limbs, arm and leg specific versions

new procedure is called autologous fat transfer (AFT).

Piatkowski and colleagues performed an RCT to compare IBR and AFT. In terms of the HRQL assessment itself, Piatkowski and colleagues administered the BREAST-Q, [27] a validated PROM, and examined the outcome at 12 months. This particular HRQL instrument measures QoL, has several domains, five of which the authors used: “Psychosocial Well-Being,” “Sexual Well-Being,” “Physical Well-Being: Chest,” “Satisfaction with Breasts,” and “Satisfaction with Outcome.”

Have Important Aspects of HRQL Been Omitted?

Depending on the condition assessed, it is important that clinical investigators remain unbiased in assessing HRQL. In other words, their assessment should be comprehensive. Inclusion of some HRQL issues and exclusion of others will bias the final results.

Some investigators use only generic HRQL (e.g., SF-36) in their studies. However, generic HRQL scales may not capture the specific differences of IBR and AFT. On the other hand, a breast-reconstruction-specific HRQL or PROM, if one exists, should be able to discriminate between the two approaches to post-mastectomy reconstruction. It is for these reasons that in surgical trials we should administer three types of questionnaires (i.e., generic, condition-specific, and utility), as suggested by Guyatt et al., to ensure that we are taking into account all aspects of the patients’ health and quality of life [28].

As mentioned, earlier breast reconstruction after mastectomy can greatly impact one’s quality of life and would require a disease-specific HRQL instrument. The investigators took this into account, ensuring they chose the correct HRQL measures to assess health status (i.e., the BREAST-Q, a condition specific HRQL instrument). However, in this article, they do not report a generic or health utility scale, measured with instruments such as the EQ-5D or the HUI-3. As we recall from our initial literature search, there were sub-studies published based off the same

RCT. We return to our search results to check these studies. Indeed, in a separate publication from Wederfoort et al. [29], they report a cost-effectiveness analysis based off the same RCT, in which they reported EQ-5D-5L and quality-adjusted life years (QALYs). Later in the chapter, we discuss the cost-analysis. However, it was not clear by reading Piatkowski et al. [7] alone whether they implemented any other HRQL measures, thus, we suggest a report and discussion on the generic HRQL outcome (i.e., the EQ-5D-5L) would have strengthened this paper.

Death is always a possibility with any surgical procedure. Although death is rare in the course of postmastectomy reconstruction if there was a difference the investigators should report it. In the Piatkowski et al. study [7], no deaths were reported. If patients were of an employable age, reporting on their ability to return to work would be an important issue for the patients. As the two approaches of postmastectomy reconstruction vary significantly in terms of the number of surgical procedures to reconstruct the breast, the days off work should have been reported.

Are the HRQL Instruments Chosen Valid, Reliable, and Responsive?

For an HRQL instrument to be chosen, it must meet three important preconditions: validity, reliability, and responsiveness to change. This ensures that it is measuring HRQL (the outcome) in a way that provides consistent, valid results. In short, the validity of an instrument describes to what degree it measures what it was developed to measure.

Reliability is assessed by conducting tests of repeatability or reproducibility of the measure under consideration. It can be classified as either having high interobserver (between different observers) reliability, high intra-observer (within the same observers) reliability, or both. These measures provide information about whether the measure being used has a tendency to change depending on who is conducting the test or when it is being conducted. Ideally, we are after the least variable inter- and intra-observer scores. This would mean that regardless of assessor and

time period, the measure under consideration is providing reliable information on which to base conclusions.

Test–retest reliability is measured by comparing scores from the same individual and then calculating a correlation between the different results. Piatkowski et al. [7] did not assess test–retest reliability. This is important because it illustrates whether results are truly reliable based on whether the calculated correlation is strong. It is important to keep in mind that a reliable test is not always a valid one, but a valid test is always reliable so it is necessary to evaluate both these qualities of an assessment tool.

Last, responsiveness has two major aspects: internal and external. Internal responsiveness entails the ability of a measure to capture the change in a participant’s outcome over a prespecified timeframe. In the case of Piatkowski et al. [7], that would be 12 months. External responsiveness reflects the extent to which change in a measure relates to a corresponding change in a reference measure of clinical or health status [30, 31]. Responsiveness, therefore, tells us whether or not an assessment tool will be able to detect a change in the outcome (HRQL) over time (12 months).

Piatkowski et al. [7] used the BREAST-Q, which has demonstrated high validity, and reliability, as supported in clinical research [32–34]. The responsiveness of the BREAST-Q has also been documented [35]. We conclude that the measures used for the measurement of HRQL by Piatkowski et al. [7] meet the guidelines of validity, reliability, and responsiveness to change.

a meaningful change, if any. If you measure the HRQL for example, at 1 week after the surgery, the patients may still be recovering from the actual surgery and no meaningful information will be obtained. Additionally, if the follow-up time is not long enough, there may not be the opportunity to observe any complications posttreatment.

In the Piatkowski et al. [7] study, however, in which the AFT approach required multiple surgical procedures to accomplish the desired reconstruction, the investigators spaced their assessments too far apart, that is, 6 and 12 months. There may have been a clinically important difference in the QoL of patients between the two techniques which may have been missed, for example, at 2 and 3 months. Multiple surgical procedures with AFT may have caused more pain, and more anxiety which would have impacted the patient’s QoL. By only measuring the HRQL at 6 and 12 months the investigators may have missed important intermediate time-period HRQL differences.

Additionally, we know that IBR have downstream complications such as implant ruptures with need for replacement. This is not the case for the AFT approach. A longer follow-up period may have been more appropriate, but we understand the difficulties such as cost and dropout issues in patients enrolled in RCTs that take a long time to complete.

The response shift, which is the patients’ adaptation to their illness with time, may alter the HRQL questionnaire scores. As the Piatkowski et al. [7] study was a RCT, we would not, however, expect that this would be an issue, as the response shift should apply to both groups.

Secondary Guides

Were HRQL Assessments Appropriately Timed to Evaluate the Effects of the Surgical Intervention?

Piatkowski et al. [7] administered the BREAST-Q at the time of study inclusion and at 6- and 12-months posttreatment. The timing of these assessments is critical since if performed too soon, or too close together, you may not measure

If There Were Trade-Offs Between Quantity and Quality of Life, Did the Investigators Perform an Economic Analysis?

The quantity versus quality of life is an important consideration because as mentioned earlier, the focus of surgical interventions in the past was prolonging life. Over the last decades,

it has become clear that more years or months of life does not necessarily translate to a “good” few years or months. Some patients may be willing to forgo longevity if this is associated with pain.

An economic evaluation may also be important from the perspective of the patient, third-party payers, and society. To do so, one needs to calculate the costs and effectiveness of the two comparative approaches and calculate an incremental cost-utility ratio (ICUR). This is explained in Chap. 24 of this book.

Piatkowski et al. reported an economic evaluation comparing the two approaches to breast reconstruction in a separate report [29]. They conducted a cost-effectiveness analysis using quality-adjusted life years (QALY) as an outcome to measure the quality of life for patients. Specifically, QALYs represent the gains from reduced mortality and morbidity in a single measure [30]. To do so, they administered a health utility measure, the EQ-5D-5L.

The EQ-5D-5L is a utility measure with 25 items to assess general health status and health-related quality of life. From this instrument, one can calculate QALYs. There are five dimensions including; mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. A health utility of 1.00 represents best possible health and a health utility of 0 or negative represents death or health worse than death [36]. There are other utility scales of course such as the HUI3, which is more a powerful scale that can discriminate between close to million different health states [37]. In the subsequent publication [29], Wederfoot et al. calculated the costs of the competing interventions (AFT and IBR) which included direct costs and productivity costs. They found that QALY and costs were higher for AFT over the first year after reconstruction. However, over the 10 and 30 year period the AFT was found to be the most effective approach since no additional surgeries are necessary in the AFT group [29].

What Were the Results?

What Was the Magnitude of the Effect on HRQL?

Investigators report the similarities or differences between scores obtained from the questionnaires, rather than whether one treatment over the other resulted in better outcomes. However, these similarities or differences need to be reported and interpreted in a meaningful way, so conclusions about HRQL of the patients following the treatments under evaluation can be made.

Piatkowski et al. [7] found that at 12 months postoperatively, the AFT group scored higher on all the BREAST-Q domains, of which the difference among three domains was statistically significant and reported as clinically relevant. These domains included “Satisfaction with Breast,” “Physical Well-Being (Chest),” and “Satisfaction with Outcome.” See Table 10.2 with results. Linear mixed-effects regression analysis also demonstrated the change in QoL over time favoring the AFT group for the domains of “Satisfaction with Breasts” (difference, 6.2; $P = 0.005$), “Physical Well-Being (Chest)” (difference, 4.6; $P = 0.001$), and “Sexual Well-Being” (difference, 5.2; $P = 0.003$).

As previously mentioned, patient-reported outcomes (PROs) have become increasingly utilized in clinical research. The questionnaires used in Piatkowski et al. (i.e., the BREAST-Q) is a powerful way to measure PROs. Interpreting the scores of these questionnaires properly is immensely important, as this is the information that is used to base the decision of whether a novel treatment will be used over the current gold-standard treatment. A way to interpret scores or responses is by using a minimal clinically important difference (MCID) that has been supported in previous literature for the surveys in question. The MCID has been defined as “the smallest difference in score in the domain of interest in which patients perceive as beneficial” [38]. Clinicians use the MCID to define the level of change an intervention creates to determine whether that change is a clinically important, or relevant change in the population [38]. As mentioned earlier,

Table 10.2 BREAST-Q scores

BREAST-Q domain	BREAST-Q score (mean, SD)									
	Study inclusion		6 months				12 months			
	AFT (n = 80)	IBR (n = 76)	AFT (n = 63)	IBR (n = 55)	Difference	P value	AFT (n = 64)	IBR (n = 68)	Difference	P value
Satisfaction with breasts	60.0 (23.4)	61.8 (22.9)	66.3 (16.6)	62.4 (16.17)	3.9	0.21	70.3 (17.8)	60.4 (17.2)	9.9	0.002
Psychosocial well-being	62.5 (17.5)	65.8 (16.0)	65.6 (16.0)	69.1 (18.1)	−3.5	0.27	69.6 (17.4)	68.3 (17.9)	1.4	0.66
Physical well-being (chest)	77.4 (14.3)	79.2 (14.3)	79.1 (14.6)	73.7 (15.9)	5.4	0.06	79.9 (14.7)	72.3 (17.0)	7.6	0.007
Sexual well-being	54.4 (16.8)	61.1 (16.6)	57.3 (18.7)	58.1 (19.0)	−0.8	0.81	61.5 (18.6)	58.6 (18.7)	2.9	0.37
Satisfaction with outcome	N/A ^a	N/A ^a	70.1 (20.1)	73.2 (20.4)	2.4	0.52	73.9 (22.4)	66.3 (19.8)	7.6	0.04

^aAdapted from Table 2 in Piatkowski et al. [7]

Piatkowski et al. [7] reported the statistically significant differences as clinically relevant as well, referencing previously determined BREAST-Q MCID estimates for the BREAST-Q. The MCID estimates have been previously defined as 4, for the domains “Satisfaction with Breasts,” “Psychosocial Well-Being,” “Sexual Well-Being,” and 3 for “Physical Well-Being (Chest)” [39]. However, there is not yet a reported MCID for “Satisfaction with Outcome.” At 12 months, the difference in means reported for “Satisfaction with Breasts” and “Physical Well-Being (Chest)” were statistically significant ($P < 0.05$) and larger than the respective MCIDs (>4 and >3), indeed making them clinically important as well. The differences in means for Psychosocial Well-Being and Sexual Well-Being were both statistically nonsignificant ($P > 0.05$) and did not meet the MCID (<4). The difference in means for “Satisfaction with Outcome” was statistically significant.

whether our patients will be helped by the results of the Piatkowski et al. [7] study, we would like to know if the patients are similar to our patient. We, therefore, look at the table that describes the demographic characteristics of their study group. We look to see if the Piatkowski et al. [7] study patients were similar in age to our patient and any other prognostic factors.

Our patient is 47 years in age and the mean patient age in the Piatkowski et al. [7] study was 49 years, (just 2 years older) which is fairly similar to ours. If we believe that their patient group (Dutch) is different from ours (North American), we may be skeptical of the generalizability of their findings. If we agree that there are no major differences in these two populations, we should accept their findings. As we do not have any evidence of major differences, we should accept their findings.

Will the Results Help Me in Caring for My Patients?

Will the Information from This Study Help Me Inform My Patients?

Surgeons know from their experience that patients who undergo similar surgical interventions can respond differently. To determine

Resolution of the Scenario

The results from the Piatkowski et al. [7] study indicate that there is important differences in the HRQL of patients who undergo AFT. Although the serious adverse events were similar between the two groups, there appeared to be twice as many adverse events in the AFT. In this scenario, these results are explained to the patient. Knowing

that the AFT provides better QoL and she does not have to deal with the downstream complications of the IBR such as deflations and repeat surgeries for capsular contracture releases, she opted to undergo the AFT approach.

Appendix

Search Strategy on Cochrane

Keywords: “breast reconstruction” AND “autologous fat transfer” AND “implant-based reconstruction” AND “randomized controlled trial” AND “quality of life”

Time horizon: last 10 years (2013–April 2024)

Search Results

1. Wederfoort JLM, Schop S, van der Broeck LCA, Hommes JE, van Kuijk SMJ, Timmermans F, Smit JM, Heuts EM, de Wit T, van der Hulst RRWJ, Piatkowski AA. Superior sensibility after full breast reconstruction with autologous fat transfer. *Plast Reconstr Surg.* 2024;153(2):316–23. <https://doi.org/10.1097/PRS.00000000000010619>.
2. Piatkowski AA, Wederfoort JLM, Hommes JE, Schop SSJ, Krastev TK, van Kuijk SMJ, van der Hulst RRWJ; Breast Reconstruction With External Preexpansion & Autologous Fat Transfer vs Standard Therapy (BREAST) Trial Investigators. Effect of total breast reconstruction with autologous fat transfer using an expansion device vs implants on quality of life among patients with breast cancer: a randomized clinical trial. *JAMA Surg.* 2023;158(5):456–464. <https://doi.org/10.1001/jamasurg.2022.7625>.
3. Wederfoort JLM, Voeten N, Smeins NNPM, Hommes JE, Essers BAB, van der Hulst RRWJ, Piatkowski A; BREAST-trial Investigators. A Dutch randomized controlled study shows autologous fat transfer with pre-expansion is more cost-effective in the long term than implants. *J Plast Reconstr Aesthetic Surg.* 2023;84:398–412.

References

1. World Health Organization. WHOQOL: measuring quality of life [Internet]. Geneva; 2024 [cited 2024 Apr 26]. Available from: <http://www.who.int/healthinfo/survey/whoqol-qualityoflife/en/>.
2. Centers for Disease Control and Prevention. Health-Related Quality of Life (HRQOL) [Internet]. 2022 [cited 2024 Apr 26]. Available from: <https://archive.cdc.gov/#/details?url=https://www.cdc.gov/hrqol/index.htm>.
3. Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis.* 1985;38:27–36.
4. Wood-Dauphinee S. Quality of life assessment: recent trends in surgery. *Can J Surg.* 1996;39:368–72.
5. Wood-Dauphinee S. Assessing quality of life in clinical research: from where have we come and where are we going? *J Clin Epidemiol.* 1999;52:355–63.
6. Waltho D, Kaur MN, Haynes RB, Farrokhyar F, Thoma A. Users' guide to the surgical literature: how to perform a high-quality literature search. *Can J Surg.* 2015;58(5):349–58.
7. Piatkowski AA, Wederfoort JLM, Hommes JE, Schop SSJ, Krastev TK, van Kuijk SMJ, van der Hulst RRWJ, Breast Reconstruction With External Preexpansion & Autologous Fat Transfer vs Standard Therapy (BREAST) Trial Investigators. Effect of total breast reconstruction with autologous fat transfer using an expansion device vs implants on quality of life among patients with breast cancer: a randomized clinical trial. *JAMA Surg.* 2023;158(5):456–64.
8. US Food and Drug Administration. Clinical outcome assessment assessment: frequently asked questions [Internet]. 2020 [cited 2024 Apr 26]. Available from: <https://www.fda.gov/about-fda/clinical-outcome-assessment-coa-frequently-asked-questions#COADefinition>.
9. Streiner DL, Norman RG, Cairney J. Health measurement scales: a practical guide to their development and use. 5th ed. Oxford: Oxford University Press; 2014.
10. McDowell I. Measuring health: a guide to rating scales and questionnaires. 3rd ed. Oxford: Oxford University Press; 2006.
11. Deshpande PR, Rajan S, Sudeepthi Nazir CPA. Patient-reported outcomes: a new era in clinical research. *Perspect Clin Res.* 2011;2(4):137–44.
12. Hou ZK, Liu FB, Fang JQ, Li XY, Li LJ, Lin CH. Advising on preferred reporting items for patient-reported outcome instrument development: the PRIPROID. *Chin J Integr Med.* 2013;19(3):172–81.
13. Litwin MS. Chapter 13: Health-related quality of life. In: Penson DF, Wei JT, editors. *Clinical research for surgeons*. Totowa, NJ: Humma Press; 1998.
14. Wang K, Yee C, Tam S, Drost L, Chan S, Zaki P, Rico V, Ariello K, Dasios M, Lam H, DeAngelis C, Chow E. Prevalence of pain in patients with breast

- cancer post-treatment: a systematic review. *Breast*. 2018;42:113–27.
15. DiSipio T, Rye S, Newman B, Hayes S. Incidence of unilateral arm lymphoedema after breast cancer: a systematic review and meta-analysis. *Lancet Oncol*. 2013;14(6):500–15.
 16. Cordeiro PG. Breast reconstruction after surgery for breast cancer. *N Engl J Med*. 2008;359(15):1590–601.
 17. Dean C, Chetty U, Forrest APM. Effects of immediate breast reconstruction on psychosocial morbidity after mastectomy. *Lancet*. 1983;1(8322):459–62.
 18. Eltahir Y, Werners LLCH, Dreise MM, van Emmichoven IAZ, Jansen L, Werker PMN, de Bock GH. Quality-of-life outcomes between mastectomy alone and breast reconstruction: comparison of patient-reported BREAST-Q and other health-related quality-of-life measures. *Plast Reconstr Surg*. 2013;132(2):201e–9e.
 19. Elder EE, Brandberg Y, Björklund T, Rylander R, Lagergren J, Jurell G, Wickman M, Sandelin K. Quality of life and patient satisfaction in breast cancer patients after immediate breast reconstruction: a prospective study. *Breast*. 2005;14(3):201–8.
 20. Atisha DM, Rushing CN, Samsa GP, Locklear TD, Cox CE, Shelley Hwang E, Zenn MR, Pusic AL, Abernethy AP. A national snapshot of satisfaction with breast cancer procedures. *Ann Surg Oncol*. 2015;22(2):361–9.
 21. Dean NR, Crittenden T. A five year experience of measuring clinical effectiveness in a breast reconstruction service using the BREAST-Q patient reported outcomes measure: a cohort study. *J Plast Reconstr Aesthet Surg*. 2016;69(11):1469–77.
 22. Khajuria A, Prokopenko M, Greenfield M, Smith O, Pusic AL, Mosahebi A. A meta-analysis of clinical, patient-reported outcomes and cost of DIEP versus implant-based breast reconstruction. *Plast Reconstr Surg Glob Open*. 2019;7(10):e2486.
 23. Alderman AK, Wilkins EG, Kim HM, Lowery JC. Complications in postmastectomy breast reconstruction: two-year results of the Michigan Breast Reconstruction Outcome Study. *Plast Reconstr Surg*. 2002;109(7):2265–74.
 24. Bircoll M. Cosmetic breast augmentation utilizing autologous fat and liposuction techniques. *Plast Reconstr Surg*. 1987;79(2):267–71.
 25. Spear SL, Coles CN, Leung BK, Gitlin M, Parekh M, Macarios D. The safety, effectiveness, and efficiency of autologous fat grafting in breast surgery. *Plast Reconstr Surg Glob Open*. 2016;4(8):e827.
 26. Khouri RK, Rigotti G, Khouri RK Jr, Cardoso E, Marchi A, Rotemberg SC, Baker TJ, Biggs TM. Tissue-engineered breast reconstruction with BRAVA-assisted fat grafting: a 7-year, 488-patient, multicenter experience. *Plast Reconstr Surg*. 2015;135(3):643–58.
 27. BREAST-Q, version1.0. Memorial Sloan Kettering Cancer Center. <https://qportfolio.org/wp-content/uploads/2023/01/BREAST-Q-BREAST-CANCER-USER-GUIDE.pdf>. Accessed April 26, 2024.
 28. Guyatt GH, Naylor CD, Juniper E. Users' guides to the medical literature XII. How to use articles about health-related quality of life. *JAMA*. 1997;277:1232–7.
 29. Wederfoort JLM, Voeten N, Smeins NNPM, Hommes JE, Essers BAB, van der Hulst RRWJ, Piatkowski A, BREAST-trial Investigators. A Dutch randomized controlled study shows autologous fat transfer with pre-expansion is more cost-effective in the long term than implants. *J Plast Reconstr Aesthet Surg*. 2023;84:398–412.
 30. Thoma A, Cornacchi SD, Lovrics PJ, Goldsmith CH. Users' guide to the surgical literature: how to assess an article on health-related quality of life. *Can J Surg*. 2008;51(3):215–24.
 31. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol*. 2000;53:459–68.
 32. Pusic AL, Klassen AF, Scott AM, Klok JA, Cordeiro PG, Cano SJ. Development of a new patient-reported outcome measure for breast surgery: the BREAST-Q. *Plast Reconstr Surg*. 2009;124:345–53.
 33. Cano SJ, Klassen AF, Scott AM, Cordeiro PG, Pusic AL. The BREAST-Q: further validation in independent clinical samples. *Plast Reconstr Surg*. 2012;129:293–302.
 34. Kaur MN, Chan S, Bordeleau L, Zhong T, Tsangaris E, Pusic AL, Cano SJ, Klassen AF. Re-examining content validity of the BREAST-Q more than a decade later to determine relevance and comprehensiveness. *J Patient Rep Outcomes*. 2023;7(37).
 35. Pusic AL, Matros E, Fine N, Buchel E, Gordillo GM, Hamill JB, Kim HM, Qi J, Albornoz C, Klassen AF, Wilkins EG. Patient-reported outcomes 1 year after immediate breast reconstruction: results of the mastectomy reconstruction outcomes consortium study. *J Clin Oncol*. 2017;35(22):2499–506.
 36. EuroQoL Group. EQ-5D-5L [Internet]. 2024 [cited 2024 Apr 26]. Available from: <https://euroqol.org/eq-5d-instruments/eq-5d-5l-about/>.
 37. Horsman J, Furlong W, Feeny D, Torrance G. The health utilities index (HUI): concepts, measurement properties and applications. *Health Qual Life Outcomes*. 2003;1(54):1–13.
 38. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10:407–15.
 39. Voineskos SH, Klassen AF, Cano SJ, Pusic AL, Gibbons CJ. Giving meaning to differences in BREAST-Q scores: minimal important difference for breast reconstruction patients. *Plast Reconstr Surg*. 2020;145(1):11e–20e.

Randomized Controlled Trial Comparing Surgical Interventions

11

Lily J. Park, Marko Simunovic, Mohit Bhandari,
and Pablo E. Serrano

Abbreviations

ARR	Absolute Risk Reduction
BMI	Body Mass Index
CDC	Centres for Disease Control
CHF	Congestive Heart Failure
CI	Confidence Interval
COPD	Chronic Obstructive Pulmonary Disease
DVT	Deep Vein thrombosis
HbA1C	Hemoglobin A1C
HR	Hazard Ratio
IQR	Interquartile Range
MCID	Minimal Clinically Important Difference
MRSA	Methicillin-Resistant <i>Staphylococcus Aureus</i>
NNT	Number Needed to Treat
PE	Pulmonary Embolus
PICOT	Population, Intervention, Comparative interventions, Outcome, Time period
RCT	Randomized Controlled Trial
RR	Relative Risk
RRR	Relative Risk Reduction

SCIP	Standard Surgical Care Improvement
TKA	Total Knee Arthroplasty
USD	United States Dollars

Introduction

Despite years of training and schooling, clinicians routinely encounter difficult clinical scenarios that are atypical of their everyday practice. Historically, physicians made therapeutic decisions based on anecdotal reports and personal experience. More recently, guidelines endorse reduced variability among patient care and highlight the importance of applying evidence from clinical research [1, 2]. Therefore, to answer questions from unfamiliar clinical scenarios, physicians rely on high-quality research evidence—or in other terms, evidence-based practice. With high-quality research being integral to the practice of evidence-based medicine, it is no surprise that the National Institutes of Health, in the United States, invests nearly \$37.7

L. J. Park
Division of General Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada

Department of Health Research Methods, Evidence, and
Impact, McMaster University, Hamilton, ON, Canada
e-mail: lily.park@medportal.ca

M. Simunovic
Division of General Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: simunovi@hhsc.ca

M. Bhandari
Department of Surgery, Division of Orthopaedic
Surgery, McMaster University,
Hamilton, ON, Canada
e-mail: bhandarm@mcmaster.ca

P. E. Serrano (✉)
Division of General Surgery, Department of Surgery
& Department of Health Research Methods, Evidence
& Impact, McMaster University,
Hamilton, ON, Canada
e-mail: serrano@mcmaster.ca

billion annually toward funding medical research projects [3].

In the hierarchy of evidence based medicine, randomized controlled trials (RCTs) are emphasized as the optimal primary study design for reducing bias and capturing the truest estimates of treatment effect. Yet, not all RCTs are equal in terms of validity. Simply stating that a study was randomized does not ensure it is a high-quality study. The purpose of this chapter is to introduce strategies for clinicians to use while evaluating the evidence, with a particular focus on RCTs comparing surgical interventions. Arguably, conducting an RCT in a surgical setting poses some unique challenges, such as with blinding and the surgical learning curve, that if not properly accounted for, can lead to potential confounders. We will illustrate how physicians may apply these strategies and identify potential confounders with the help of a clinical scenario.

Scenario

A healthy 70-year-old woman comes into your clinic complaining of a palpable bulge along her previous laparotomy scar. A year prior to her presentation, she had undergone a ventral hernia repair with synthetic mesh concomitantly with an elective cholecystectomy for biliary colic. On review of her imaging studies and clinical exam, you conclude that she has recurrence of her hernia, which is currently soft, reducible, without signs of incarceration or strangulation. You discuss with her the nature of her problem, as well as the risks and benefits of further surgery or a watch-and-wait approach. Your patient mentions that she heard about biologic mesh and asks you “If I had my initial hernia fixed with a different mesh, would it still have failed?” Unsure initially of how to answer this question, you decide to review the literature and assure her that at her next clinical appointment, you will provide her with the best possible answer.

Literature Search

You begin your literature search by creating a well-defined research question that encompasses several aspects of the clinical scenario. Using the

PICO(T) format [4], which incorporates information about the patient population (P), the intervention (I), comparative interventions (C), the outcome (O), and time period (T) (see Chap. 3), you generate the following research question: In patients with a ventral hernia undergoing a concomitant clean-contaminated procedure, does fixation with a synthetic mesh lead to higher recurrence rates compared to biologic mesh? Using Medline, the National Library of Medicine’s PubMed database [5], you enter “ventral hernia” AND “synthetic mesh” AND “recurrence” in the search field (Appendix 1). You limit your search to English articles, published in the last 3 years, randomized controlled trials, and human adult subjects. The search yields eight articles [6–13]. You choose to review the most recent article entitled, “Biologic vs Synthetic Mesh for Single-Stage Repair of Contaminated Ventral Hernias: A Randomized Clinical Trial” [6], which compares two methods of ventral hernia repair and seems to address your clinical question.

Summary of the Appraised Article

The study by Rosen et al. [6] was a multicenter, randomized controlled trial comparing synthetic versus biologic mesh among patients who underwent single-stage repair of clean-contaminated and contaminated ventral hernias. The trial included 253 patients across 5 clinical centers. Following randomization, 127 patients were assigned to receive a biologic intervention and 126 patients to receive a synthetic intervention. The primary outcome of this study was hernia recurrence within 2 years as a composite of clinical exam findings, a blinded review of radiographic imaging, or patient-reported outcome of a bulge on the hernia recurrence inventory. Other important outcomes included postoperative adverse events, quality of life, cost, and mesh safety. Mesh safety was assessed by the need for a procedural intervention within 2 years including the need for opening of the surgical site, wound debridement, suture excision, percutaneous drainage, or partial or complete mesh removal. The follow-up rate after 2 years was 92%.

Synthetic mesh significantly reduced the risk of hernia recurrence compared to biologic mesh (5.6% vs. 20.5%, hazard ratio [HR] 0.31, 95% confidence interval [CI] 0.23–0.42, $p < 0.001$). The median [interquartile range (IQR)] 30-day direct hospital costs were significantly greater in the biologic group compared to the synthetic group [\$44,936 (\$35,877–\$52,656) vs \$17,289 (\$14,643–\$22,901), $p < 0.001$, in United States Dollars (USD)]. There were no significant differences in quality of life or the need for a procedural intervention between the two groups.

Evaluating a Randomized Controlled Trial

When reading a research article, it is important to evaluate whether the study design and execution were appropriate in answering the research question with reliable results. Notably, three questions should be asked: Are the results valid?, What are the results?, And are the results applicable to my practice (see Box 11.1) [14]?

Box 11.1 Important points to consider when evaluating a surgical RCT

Are the results valid?

- Was the learning curve of a procedural skill taken into consideration?
- How was randomization performed?
- How was allocation concealment performed?
- Who was blinded?
- Were the patient groups similar to one another?
- How were patients' data analyzed?
- Were treatments standardized and all patients accounted for?

What are the results?

- What impact did the treatment have?
- How precise were the results?

Are the results applicable to my practice?

- Do the results apply to my patient population?
 - How clinically relevant are the results?
 - How do these results impact me?
- This list was modified from Ref. [14].

Are the Results Valid?

Learning Curve and Expertise-Based Randomized Controlled Trials

The surgical learning curve is an important factor to consider when evaluating surgical RCTs. This learning curve refers to the belief that a surgeon's proficiency, efficiency, and expectant outcomes of a procedure will improve with experience [15, 16]. Simply speaking, the more cases a surgeon has completed, the better they will become with the technical aspects of the procedure. Likewise, with experience, surgeons will have a better understanding of the necessary adjunctive care like medications, patient selection, and various other pre- and postoperative care regimens that can optimize their outcomes. If an RCT does not consider (or mention) a learning curve, especially when studying a "novel" intervention, the results may be biased in favor of the traditional, or more commonly performed, intervention. One solution to this challenge is the "expertise-based RCT." Unlike the traditional RCT where patients are randomized to intervention A or B, in expertise based RCTs, the patient would be randomized directly to a surgeon who is an expert in intervention A or a surgeon who is an expert in intervention B. In this way, surgeons would only deliver the intervention in which they are an expert [17]. However, expertise-based trials are not without fault as determining expertise, achieving adequate recruitment, and the context-specific nature of such studies prove challenging [18, 19].

In the trial by Rosen et al. [6], there were eight participating surgeons who were all fellowship-trained in advanced abdominal wall reconstruction. With this in mind, we can assume that participating surgeons had sufficient expertise in performing either intervention.

Patient Randomization Methods

Randomization is a technique that assigns patients to either the treatment or control arm of the study in a manner that is entirely by chance and without taking into consideration patient or researcher preference [20]. The randomization schedule can often be created by computer-generated sequences. The purpose of randomization is to produce groups that are similar to one another in terms of both known and unknown characteristics that may otherwise influence the study outcome. When randomization is executed correctly, we eliminate the bias in choice of treatment assignment, such that we can assume that any baseline differences between groups are due to chance. Without randomization, researcher, patient, and physician bias may influence experimental outcomes and alter study results.

Randomization methods can be fixed or adaptive [21, 22]. Fixed randomization follows a predetermined randomization process that is constant throughout the trial. When randomization is fixed, there is usually a predetermined allocation ratio, which represents the odds of randomization to each treatment group [23]. For instance, if the allocation ratio is 1:1, the probability of being randomized to intervention A or B is 50%. Unequal allocation ratios (e.g., 2:1) can also be used, although this is less commonly done due to reduced efficiency for power. Fixed randomization is most commonly used as it is simple and less computationally intensive. However, in small samples, fixed randomization can lead to chance imbalances in important baseline characteristics between treatment groups. Thus, adaptive randomization have become increasingly popular, where randomization is modified during the trial based on accumulating data.

Common fixed randomization methods include simple, restricted, and stratified randomization. In simple or unrestricted randomization, every participant has an equal chance of being randomized to each treatment group. However, this can lead to imbalances in the number of patients in each of the groups. Restricted or blocked randomization use block sizes to ensure balance in the number of participants in each treatment group, at the end of the trial. For instance, if the block size is four, then every four consecutive participants will have an equal chance of being assigned to each treatment

group. This block size can be fixed (e.g., 4 every time) or random (e.g., could be block size of 4 or 6 at any given time). Stratified randomization helps to ensure balance of specific predetermined characteristics that are considered important for the study. Commonly, randomization is stratified by known confounding factors. So if, for example, randomization is stratified by disease severity, participants would be grouped into their corresponding strata (e.g., mild, moderate, severe) with subsequent randomization performed within each stratum, to ensure that all groups are balanced in these key characteristics that could impact the results.

Common adaptive randomization methods include minimization and response adaptive randomization. In response adaptive randomization, the allocation ratio changes over time to assign more patients to the group that is having better outcomes [24]. Although the advantage would be a reduction in the number of patients that receive inferior treatment, the adjustment of treatment assignments based on interim results could introduce selection bias and compromise the validity of the study [24]. Minimization is another adaptive randomization method where the group assignment of each participant is dependent on the balance in prognostic factors that is present in the trial at the time of treatment allocation [25–27]. Minimization has been shown to provide more balanced groups across a greater number of prognostic factors compared to the aforementioned fixed randomization methods [25–27]. However, considering a non-random component of treatment allocation that is introduced, it is particularly important that the minimisation algorithm and group allocation is not accessible by any individual who may introduce bias [25–27].

The Rosen et al. trial implemented a fixed randomization method with 1:1 allocation to the synthetic and biologic mesh groups. They used a central concealed randomization scheme with random block sizes and stratification by medical centre and centers for disease control (CDC) wound class [6].

Allocation Concealment

Similarly, allocation concealment is crucial to limiting selection bias by concealing which treat-

ment arm each prospective study patient will be assigned to [20]. In essence, it is a strategy to protect the integrity of the randomization sequence. An example of allocation concealment is the use of a central call-in center or computer program, which will reveal the treatment arm a patient has been randomized to only after enrolment. As such, physicians cannot predict which treatment the patient will receive prior to enrolment and randomization. Despite proper randomization, failure to adequately control allocation concealment may introduce bias and influence study results. For example, if investigators opened unsealed assignment envelopes and channelled participants with a better prognosis to the experimental group, this would lead to larger treatment effects [28].

In the Rosen et al. trial [29], the randomization sequence was generated centrally by the study statistician and local research coordinators at each site used a centralised computer system to execute allocation intraoperatively, immediately prior to mesh deployment. Thus, they utilized a methodologically robust approach for randomization and allocation concealment.

Were Patients, Surgeons, and Researchers Aware of Group Assignment (Blinding)?

Within RCTs, blinding refers to the precautions taken to prevent the patient, surgeon, or researchers from knowing a participant's group assignment throughout the trial [30]. Blinding minimizes bias in intervention implementation, outcomes assessment, results analysis, and patient dropout. Take, for example, the placebo effect, a scenario where a patient's belief about their treatment influences their outcomes [31]. In a meta-analysis comparing osteoarthritis patients receiving a placebo agent to an untreated control group, patients in the placebo group experienced significantly more pain and stiffness relief than the control group [32], even though neither groups received active treatment. If patients are aware of their group allocation they may alter their answers on quality assessments and subconsciously put forth more of an effort in their rehabilitation, exaggerating the experimental treatment effect. Similarly, surgeons may favor one intervention over another

and may unintentionally be more precise during its implementation, ultimately overestimating study results [33, 34]. Unfortunately, the very nature of surgery makes it nearly impossible for surgeons to remain blinded during intervention implementation. Finally, the research personnel in charge of assessing the outcomes, if not blinded, may alter study results. If, for example, the researchers rounded outcomes from the treatment arm of the study upwards compared to the control group, or if preferential treatment were given to either group during rehabilitation or assessment, study results could be distorted [35, 36].

In the Rosen et al. trial [6], patients and radiographic assessors for recurrence were blinded until the end of the study, whereas surgeons were not blinded. Unfortunately, as there is no real solution to surgeon blinding, it is difficult to comment on whether or not differential care was provided to either care group. However, the study attempted to reduce chances of differential care as much as possible by withholding randomization until time of mesh deployment. This, along with the incorporation of multiple surgeons may have helped to balance any differences in procedure implementation between the two groups.

Were the Patients in Each Group Similar?

At the onset of the trial, it is imperative that the experimental and control groups be relatively homogenous. In other words, the more alike the two groups are to one another prior to trial commencement the less likely that potential prognostic factors can influence study results. In effect, we would be able to detect the true effects of the therapeutic intervention. Most commonly, studies present patient baseline demographic information and known prognostic variables—often presented in a table. The Rosen et al. trial [6] included adult patients with contaminated wounds (CDC class II or III), undergoing elective single-stage ventral hernia repairs. The baseline patient characteristics presented in Table 11.1 demonstrates a good balance between the two groups. This lends support to having developed and executed an appropriate randomization strategy.

Table 11.1 Patient baseline characteristics [6]

	Biologic mesh (<i>n</i> = 127), <i>n</i> (%)	Synthetic mesh (<i>n</i> = 126), <i>n</i> (%)
Age, years in median (IQR)	63.7 (55–70)	63.7 (55–69)
Sex, male	56 (44)	61 (48)
BMI in median (IQR)	32.9 (29–35)	32.3 (28–37)
Obese (BMI >30)	81 (64)	73 (58)
History of ever smoking	81 (64)	71 (56)
Hypertension	80 (63)	73 (58)
Diabetes	36 (28)	29 (23)
COPD	21 (16)	12 (10)
Dyspnea	8 (6)	8 (6)
CHF	6 (5)	3 (2)
Prior DVT	18 (14)	14 (11)
Prior PE	8 (6)	15 (12)
Crohn's disease	14 (11)	13 (10)
Ulcerative colitis	12 (9)	9 (7)
History of cancer	48 (38)	46 (37)
Previous radiotherapy	15 (12)	14 (11)
Serum HbA1c, median (IQR)	5.7 (5.4–6.1)	5.7 (5.5–6.3)
Recurrent incisional hernia	67 (53)	59 (47)
Stoma present	82 (65)	77 (61)
Recurrent parastomal hernia	29 (23)	30 (24)
Prior mesh infection	9 (7)	7 (6)
History of abdominal wall infection	24 (19)	25 (20)
Nonhealing abdominal wound	16 (13)	16 (13)
Prior MRSA infection	9 (7)	10 (8)

IQR interquartile range, *BMI* body mass index, *COPD* chronic obstructive pulmonary disease, *CHF* congestive heart failure, *DVT* deep vein thrombosis, *HbA1C* hemoglobin A1C, *MRSA* methicillin-resistant *Staphylococcus aureus*, *PE* pulmonary embolus

All *P* values >0.05

Intention-to-Treat Analysis

Once the trial has begun, it is important to consider how the analysis was carried out. Consider the hypothetical example of an RCT comparing the rates of surgical site infection in 200 patients who underwent total knee arthroplasty. In this study, 100 patients are assigned to receive intra-operative local vancomycin powder to their wound plus preoperative ancef (experimental group), while the remaining 100 only receive the preoperative ancef for prophylaxis (control group). The surgeon then decides that 10 patients in the ancef-only group would benefit from the local vancomycin powder due to intra-operative complications. If 10 of the 100 patients in the initial experimental group develop an infection, plus 5 of the 10 patients that switched group's perioperatively go on to develop an infection, then the event rate would be 14%;

however, the rate in the control group would be 6% (Fig. 11.1). Analysis of this new grouping, irrespective of initial group assignment, would represent per-protocol analysis. In this case, per-protocol analysis would demonstrate a spurious reduction in infection rates for the control group. Intention-to-treat analysis eliminates this potential bias by analyzing patients in an RCT according to the treatment arm they were originally assigned to, irrespective of the treatment they received [37, 38]. In our example, if patients were analyzed with an intention-to-treat model the event rates would be 10/100 for both groups. Incorporating this method of analysis into a study is a conservative way to reduce bias that may arise from participant attrition or crossover. In the trial by Rosen et al. [6], the authors state that an intention-to-treat analysis was employed.

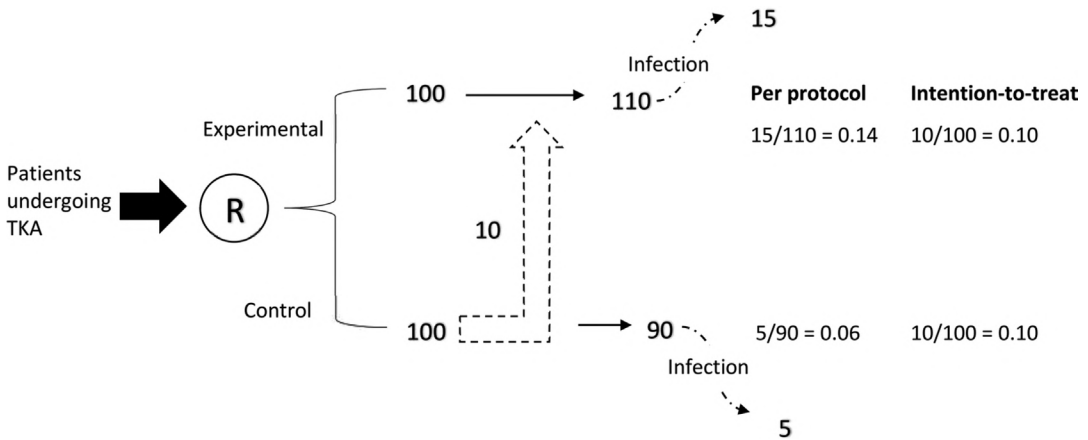


Fig. 11.1 Bias introduced when using a per-protocol analysis versus an intention-to-treat analysis

Per-protocol analysis includes all those patients who received vancomycin, irrespective of their initial assignment. With this analysis, the event rate in the experimental group is more than double that of the control group. The intention-to-treat analysis analyzes patients in their original groups irrespective of the treatment they actually received. Under this model, we see that the event rates are equal. R = randomization; TKA = total knee arthroplasty.

Treatment Standardization

Standardization of interventions is important in surgical trials to ensure treatment effects are not biased due to differential care between groups outside of the main surgical intervention (i.e., preoperative antibiotics, perioperative care, postoperative thromboprophylaxis, postoperative weight-bearing status, and rehabilitation). For example, differences in postoperative care could introduce bias if one group were given extra physiotherapy sessions. Standardization tries to address any bias that may be introduced in studies with multiple components by attempting to keep things as consistent as possible [39]. In the Rosen et al. trial [6], the Standard Surgical Care Improvement (SCIP) protocol was applied to all procedures [40]. Deep vein thrombus (DVT) prophylaxis and preoperative antibiotics were administered according to institutional protocols. The authors provide supplemental materials that detail the standardized surgical procedure in both

groups. They made extensive efforts to standardize both procedures as demonstrated by the disclosure of very specific details from the start to end of surgery including dissection, intraoperative pulse lavage antibiotic irrigation, mesh placement, drain insertion, wound closure, and drain management. Therefore, it is safe to assume that the necessary steps were taken to address any potential variability.

Sample Size Calculation and Follow-Up

A predetermined sample size calculation is integral to the conduct of an RCT (see Chap. 29). A study with too few participants may fail to adequately answer the primary study question due to being statistically underpowered. Conversely, a study should not simply recruit an excessive number of patients as this would overpower the study and may result in findings of statistical significance that are not actually clinically important [41]. As such, RCTs should have a sample size calculation based upon a determined minimal clinically important difference (MCID) and desired study power [41]. Sample size calculations generate the sample size needed to detect the minimally clinically important difference in the outcomes (usually the primary outcome) used in the calculation. However, subgroup analysis is not included in this calculation.

Another caveat in any study is participant attrition. Patients that are lost to follow-up threaten the validity of the study since we do not

know their outcomes, whether or not they died, or if there are important differences in baseline characteristics between those lost to follow-up and the remaining study group that would impact the study results [42]. While there is not any hard and fast cutoff value at which attrition related bias becomes apparent, it is generally accepted that 5% loss to follow-up is of little worry, a loss of 20% should raise concern, and a loss between 5% and 20% may still produce bias but to a lesser degree [43]. To maintain the predetermined study power and avoid skewing study results, it is important to consider the anticipated loss to follow-up when calculating the minimum sample size required for the trial in the planning phase of the study [44] (see Chap. 29 for further explanation). In the trial by Rosen et al. [6], the sample size calculation found that enrolment of 201 patients would give the trial a study power of 92% to detect a significant difference in hernia recurrence risk of 29% for biologic mesh and 9% for synthetic mesh. Assuming a maximum 20% loss to follow-up, 253 patients were necessary to adequately power the study. The authors were able to recruit 253 patients, among which 233 completed the 24-month follow-up.

What Are the Results?

What Was the Impact from the Treatment?

Now that the study has been conducted, we want to know what the results mean and their clinical importance and statistical significance.

Some common terminology used to report data are relative risk (RR), relative risk reduction (RRR), absolute risk reduction (ARR), and the number

needed to treat (NNT) (Box 11.2). The RR describes the probability of an event occurring in one group of people versus another [45, 46] (see Chap. 6 for further explanation). In our study, the RR would tell us the probability of recurrence within 24 months in patients receiving a synthetic mesh versus patients receiving biologic mesh. The RRR measures how much risk is reduced in the experimental group versus the control group [47]. Consider a hypothetical situation comparing stroke rates in hypertensive patients receiving an intensive treatment regimen (experimental) versus the standard of care (control). If 10% of the control group experienced a stroke, compared to 5% of patients in the treatment group, it can be said that the intensive treatment regimen resulted in a relative risk reduction of 50% (Table 11.2). The ARR describes the absolute difference in event rates between the control and treatment groups [47]. Using our hypothetical situation, the ARR for strokes would be 5%. The NNT refers to the number of patients that would need to be treated in order to prevent one additional adverse event from occurring and can be thought of as the inverse of the ARR [48, 49]. Referring to our hypothetical example once more, for every 20 patients treated with the intensive treatment regimen, 1 stroke would be prevented.

In the Rosen et al. trial [6], the authors report that recurrence occurred less frequently in the synthetic mesh group compared to the biologic mesh group (13% vs 20.5%, $p < 0.001$). A cox proportional hazards model was used for analysis, and results were reported as a hazard ratio (HR) with 95% confidence intervals (CI). Like RR, the HR is a measure to compare the outcome between two groups. Important distinction from RR is that HR incorporates the element of time to measure how quickly an outcome of interest (i.e.,

Table 11.2 Sample calculations from the hypothetical hypertensive drug trial

	Experimental group	Control group
Total number of patients	124	130
Number of strokes	6	13
Event rates	$x = 6/124 = 0.05$ (5%)	$y = 13/130 = 0.10$ (10%)
Relative risk: $x/y = 0.05/0.10 = 0.50$ (50%)		
Relative risk reduction: $(x - y)/y = (0.05 - 0.10)/0.10 = 0.50$ (50%)		
Absolute risk reduction: $x - y = 0.05 - 0.10 = 0.05$ (5%)		
Number needed to treat: $1/(x - y) = 1/(0.05 - 0.10) = 20$		

recurrence within 24 months) happens in one group compared to another, over time [50]. Thus, HR not only considers whether an outcome occurs, but how quickly it occurs, which can be a particularly valuable piece of information depending on the clinical scenario.

Patients in the control group were 50% more likely to experience a stroke (RR). Similarly, the experimental intervention reduced the risk of stroke by 50% (RRR). Likewise, for every 100 patients treated with the experimental intervention, there would be 5 fewer strokes (ARR). Furthermore, in order to prevent one stroke, 20 patients would need to be treated with the experimental intervention (NNT).

How Precise Were the Results?

It is impossible to know the “true” reduction in 24-month recurrence rates caused by the use of a synthetic mesh because of variables unknown to researchers that may impact the study. The best we can do is come up with a close estimate, known as the point estimate, of the true value that would lie within that ballpark. To communicate the point estimate, researchers provide a variety of values, known as the confidence interval (CI), that specifies a probability within which one can be confident the true value lies [51, 52]. By convention, a 95% CI is generally used, meaning we can be 95% certain that the “true” value lies within the interval (see Chap. 28 for more information).

Box 11.2 Equations for common statistic terminology

Relative Risk (RR)	Absolute Risk Reduction (ARR)
$RR = \frac{x}{y}$	$AAR = x - y$
Relative Risk Reduction (RRR)	Number Needed to Treat (NNT)
$RRR = \frac{(x - y)}{y}$	$NNT = \frac{1}{AAR}$

x—# of events in experimental group/total # of patients in experimental group.
y—# of events in control group/total # of patients in control group.

In the Rosen et al. trial [6], results are reported using HR, the chance of an event occurring in the treatment arm divided by the chance of the event occurring in the control arm [53], CIs, and *p* values. They report an HR of 0.31 with a 95% CI of 0.23–0.42 and a *p* value of <0.001. This means that the patients who received a synthetic mesh have reduced risk of recurrence, with the true value of this rate lying between 0.23 and 0.42. Since the confidence interval does not cross the value of no difference (i.e., HR = 1), there is a significant difference that patients who received a biologic mesh are at greater risk of recurrence. Altogether, we can conclude that the difference in event rates between the two groups are statistically significant.

What Now?

Generalizability

Before we decide to implement a particular intervention, we need to assess how relevant the information is to our patient population. In our clinical scenario, our patient is a 70-year-old woman who experienced recurrence 1 year following her ventral hernia repair for a CDC class II wound with synthetic mesh. In the Rosen et al. trial [6], they enrolled adult patients with CDC wound class II or III undergoing a single-stage ventral hernia repair. Looking at the table of patient baseline characteristics, we also see that the median age of patients was 63.7 years and that 53.8% of the

study participants were female. Altogether, we can conclude that these findings can appropriately be applied to our patient.

Clinical Relevance

It is important to consider whether the outcomes that the researchers examined are clinically relevant. For example, a study comparing two treatment modalities for rotator cuff tears that looked at “return to sport time” would be clinically useful for a 20-year-old pitcher, but less relevant for a 74-year-old patient. In the Rosen et al. trial [6], the outcomes assessed included recurrence within 24 months, safety as assessed by the need for procedural intervention at surgical site, quality of life, and cost. All these outcome measures would be relevant to both physicians and patients alike, as they are clinically important with important implications on the ultimate outcome. For instance, procedural interventions increase length of hospital stay, impose added expenses to the health care system, and expose patients to further risks and potential perioperative morbidity [54].

What Does This Mean for Me as a Healthcare Provider?

Now that we have assessed the evidence provided to us, how do we use it? For medical doctors, this may be straightforward because if a study shows that drug A is more efficacious than drug B, they can simply begin prescribing drug B. However, implementing new evidence poses potential challenges for surgeons. If a study shows that a particular intervention produces superior outcomes, how do they go about implementing it? Surgeons need to objectively critique their expertise and proficiency with a procedure and earnestly consider if they would be happy with the level of care provided. It would be unethical for a surgeon to perform a procedure they are not familiar with as patients would potentially experience a higher complication rate [55–57]. If a surgeon is not comfortable performing the procedure, they can (1) refer the patient to a colleague, (2) seek additional training, or (3) perform a different procedure after considering the evidence for it. Any of these options would be sufficient and the decision ultimately lies with each surgeon.

Resolving Our Clinical Scenario

After thoughtfully examining the information provided to us from the Rosen et al. trial [6], it is fair to conclude that synthetic mesh is superior to biologic mesh in reducing recurrence following repair of ventral hernias considered to be CDC wound class II or III. However, there are no differences with respect to quality of life or wound issues requiring procedural interventions. Although our patient in the clinical scenario experienced a hernia recurrence within a year, and may require reoperation, we can be confident in answering her that the recurrence was not due to the choice of mesh.

Appendix 1: Literature Search Strategy

“ventral hernia” AND “synthetic mesh” AND “recurrence”

- Limited to:
 - Article type: Randomized Controlled Trial.
 - Publication date: 2021-01-01 to 2024-01-01.
 - Article language: English.
 - Species: Human, Adult.

References

1. Guyatt G. Evidence-based medicine. *JAMA*. 1992;268(17):2420.
2. Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. Clinical practice guidelines we can trust [Internet]. Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, editors. Washington (DC): National Academies Press (US); 2011 [cited 2024 Jan 22]. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK209539/>.
3. National Institutes of Health (NIH). What we do: budget [Internet]. [cited 2024 Jan 22]. Available from: <https://www.nih.gov/about-nih/what-we-do/budget>.
4. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. *ACP J Club*. 1995;123(3):A12.
5. Motschall E, Falck-Ytter Y. Searching the MEDLINE literature database through PubMed: a short guide. *Oncol Res Treat*. 2005;28(10):517–22.

6. Rosen MJ, Krpata DM, Petro CC, Carbonell A, Warren J, Poulouse BK, Costanzo A, Tu C, Blatnik J, Prabhu AS. Biologic vs synthetic mesh for single-stage repair of contaminated ventral hernias: a randomized clinical trial. *JAMA Surg.* 2022;157(4):293–301.
7. Harris HW, Primus F, Young C, Carter JT, Lin M, Mukhtar RA, Yeh B, Allen IE, Freise C, Kim E, Sbitany H, Young DM, Hansen S. Preventing recurrence in clean and contaminated hernias using biologic versus synthetic mesh in ventral hernia repair: the PRICE randomized clinical trial. *Ann Surg.* 2021;273(4):648–55.
8. Miserez M, Lefering R, Famiglietti F, Mathes T, Seidel D, Sauerland S, Korolija D, Heiss M, Weber G, Agresta F, Steup WH, Śmietański M, Ribeiro R, Cuccurullo D, Catena F, Rudroff C, Rosanelli G, Schön F, Smet B, Wenger F, Saad S, Naver L, Neugebauer E. Synthetic versus biological mesh in laparoscopic and open ventral hernia repair (LAPSiS): results of a multinational, randomized, controlled, and double-blind trial. *Ann Surg.* 2021;273(1):57–65.
9. Olavarria OA, Bernardi K, Dhanani NH, Lyons NB, Harvin JA, Millas SG, Ko TC, Kao LS, Liang MK. Synthetic versus biologic mesh for complex open ventral hernia repair: a pilot randomized controlled trial. *Surg Infect.* 2021;22(5):496–503.
10. Rohit G, Kumar M, Kumar M, Surya V, Hariprasad CP, Sahu RS. Comment on: preventing recurrence in clean and contaminated hernias using biologic versus synthetic mesh in ventral hernia repair: the PRICE randomized clinical trial. *Ann Surg.* 2022;276(5):e642–3.
11. Miller BT, Krpata DM, Petro CC, Beffa LRA, Carbonell AM, Warren JA, Poulouse BK, Tu C, Prabhu AS, Rosen MJ. Biologic vs synthetic mesh for parastomal hernia repair: post hoc analysis of a multicenter randomized controlled trial. *J Am Coll Surg.* 2022;235(3):401–9.
12. Maskal SM, Thomas JD, Miller BT, Fafaj A, Zolin SJ, Montelione K, Ellis RC, Prabhu AS, Krpata DM, Beffa LRA, Costanzo A, Zheng X, Rosenblatt S, Rosen MJ, Petro CC. Open retromuscular keyhole compared with Sugarbaker mesh for parastomal hernia repair: early results of a randomized clinical trial. *Surgery.* 2024;175(3):813–21.
13. Holmdahl V, Stark B, Clay L, Gunnarsson U, Strigård K. Long-term follow-up of full-thickness skin grafting in giant incisional hernia repair: a randomised controlled trial. *Hernia.* 2022;26(2):473–9.
14. Guyatt G, Rennie D, Meade M, Cook D. User's guides to the medical literature: essentials of evidence-based clinical practice. 2nd ed. McGraw-Hill Professional; 2008. p. 380.
15. Maruthappu M, Duclos A, Lipsitz SR, Orgill D, Carty MJ. Surgical learning curves and operative efficiency: a cross-specialty observational study. *BMJ Open.* 2015;5(3):e006679.
16. Hopper AN, Jamison MH, Lewis WG. Learning curves in surgical practice. *Postgrad Med J.* 2007;83(986):777–9.
17. Devereaux PJ, Bhandari M, Clarke M, Montori VM, Cook DJ, Yusuf S, Sackett DL, Cinà CS, Walter SD, Haynes B, Schünemann HJ, Norman GR, Guyatt GH. Need for expertise based randomised controlled trials. *BMJ.* 2005;330(7482):88.
18. Cook JA. The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. *Trials.* 2009;(10):9–9.
19. Cook JA, Elders A, Boachie C, Bassinga T, Fraser C, Altman DG, Boutron I, Ramsay CR, MacLennan GS. A systematic review of the use of an expertise-based randomised controlled trial design. *Trials.* 2015;16:241.
20. Kim J, Shin W. How to do random allocation (randomization). *Clin Orthop Surg.* 2014;6(1):103–9.
21. Lim CY, In J. Randomization in clinical studies. *Korean J Anesthesiol.* 2019;72(3):221–32.
22. Friedman L, Furberg C, Demets D. Fundamentals of clinical trials: the randomization process. 5th ed. Springer.
23. Sackett D, Haynes R, Guyatt G, Tugwell P. Allocations of patients to treatments. In: *Clinical epidemiology: how to do clinical practice research.* 3rd ed; 2005. p. 85–98.
24. Korn EL, Freidlin B. Adaptive clinical trials: advantages and disadvantages of various adaptive design elements. *J Natl Cancer Inst.* 2017;109(6):dix013.
25. Treasure T, MacRae KD. Minimisation: the platinum standard for trials? *BMJ.* 1998;317(7155):362–3.
26. Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials. a review. *Control Clin Trials.* 2002;23(6):662–74.
27. Altman DG, Bland JM. Treatment allocation by minimisation. *BMJ.* 2005;330(7495):843.
28. Schulz KF. Subverting randomization in controlled trials. *JAMA J Am Med Assoc.* 1995;274(18):1456.
29. Fixation using Alternative Implants for the Treatment of Hip fractures (FAITH) Investigators. Fracture fixation in the operative management of hip fractures (FAITH): an international, multicentre, randomised controlled trial. *Lancet.* 2017;389(10078):1519–27.
30. Boutron I, Estellat C, Guittet L, Dechartres A, Sackett DL, Hróbjartsson A, Ravaud P. Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: a systematic review. *PLoS Med.* 2006;3(10):e425.
31. Kaptchuk TJ. Powerful placebo: the dark side of the randomised controlled trial. *Lancet.* 1998;351(9117):1722–5.
32. Zhang W, Robertson J, Jones AC, Dieppe PA, Doherty M. The placebo effect and its determinants in osteoarthritis: meta-analysis of randomised controlled trials. *Ann Rheum Dis.* 2008;67(12):1716–23.
33. Van der Linden W. Randomized controlled trials on the effectiveness of preoperative biliary decompression. *World J Surg.* 1988;12(1):136–7.
34. Prescott C, Gillespie G, Russell K, Colthart R, Shepherd R. Factors that limit the quality, number and progress of randomised controlled trials. *Health*

- Technol Assess [Internet]. 1999;3(20) <https://doi.org/10.3310/hta3200>.
35. Karanickolas PJ, Farrokhyar F, Bhandari M. Practical tips for surgical research: blinding: who, what, when, why, how? *Can J Surg*. 2010;53(5):345–8.
 36. Miller LE, Stewart ME. The blind leading the blind: use and misuse of blinding in randomized controlled trials. *Contemp Clin Trials*. 2011;32(2):240–3.
 37. Herman A, Botser IB, Tenenbaum S, Chechick A. Intention-to-treat analysis and accounting for missing data in orthopaedic randomized clinical trials. *J Bone Joint Surg Am*. 2009;91(9):2137–43.
 38. Montori VM, Guyatt GH. Intention-to-treat principle. *CMAJ*. 2001;165(10):1339–41.
 39. Blencowe NS, Cook JA, Pinkney T, Rogers C, Reeves BC, Blazeby JM. Delivering successful randomized controlled trials in surgery: methods to optimize collaboration and study design. *Clin Trials*. 2017;14(2):211–8.
 40. Bratzler DW, Hunt DR. The surgical infection prevention and surgical care improvement projects: national initiatives to improve outcomes for patients having surgery. *Clin Infect Dis*. 2006;43(3):322–30.
 41. Zhong B. How to calculate sample size in randomized controlled trial? *J Thorac Dis*. 2009;1(1):51–4.
 42. Dumville JC, Torgerson DJ, Hewitt CE. Reporting attrition in randomised controlled trials. *BMJ*. 2006;332(7547):969–71.
 43. Fergusson D, Aaron SD, Guyatt G, Hébert P. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. *BMJ*. 2002;325(7365):652–4.
 44. Maggard MA, O'Connell JB, Liu JH, Etzioni DA, Ko CY. Sample size calculations in surgery: are they done correctly? *Surgery*. 2003;134(2):275–9.
 45. Yelland LN, Salter AB, Ryan P. Relative risk estimation in randomized controlled trials: a comparison of methods for independent observations. *Int J Biostat*. 2011;7(1):1–31.
 46. Siström CL, Garvan CW. Proportions, odds, and risk. *Radiology*. 2004;230(1):12–9.
 47. Irwig L, Irwig J, Trevena L, Sweet M. Smart health choices: making a sense of health advice [Internet]. London: Hammersmith Press; 2008 [cited 2024 Jan 22]. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK63638/>.
 48. Haas M, Schneider M, Vavrek D. Illustrating risk difference and number needed to treat from a randomized controlled trial of spinal manipulation for cervicogenic headache. *Chiropr Osteopat*. 2010;18:9–9.
 49. Bender R. Calculating confidence intervals for the number needed to treat. *Control Clin Trials*. 2001;22(2):102–10.
 50. George A, Stead TS, Ganti L. What's the risk: differentiating risk ratios, odds ratios, and hazard ratios? *Cureus* [Internet]. 2020 [cited 2024 Jan 22];12(8). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7515812/>.
 51. Flechner L, Tseng TY. Understanding results: P-values, confidence intervals, and number need to treat. *Indian J Urol*. 2011;27(4):532–5.
 52. Sedgwick P. Randomised controlled trials: inferring significance of treatment effects based on confidence intervals. *BMJ*. 2014;349(aug18 2):g5196–g5196.
 53. Trinquart L, Attiche N, Bafeta A, Porcher R, Ravaut P. Uncertainty in treatment rankings: reanalysis of network meta-analyses of randomized trials. *Ann Intern Med*. 2016;164(10):666–73.
 54. Centres for Disease Control and Prevention (CDC). Surgical site infection event [Internet]. 2023. Available from: <https://www.cdc.gov/nhsn/pdfs/pscmanual/9pscscscurrent.pdf>.
 55. Hammond J, Queale W, Kim T, McFarland E. Surgeon experience and clinical and economic outcomes for shoulder arthroplasty. *J Bone Joint Surg Am Vol*. 2003;85(12):2318–24.
 56. Jain N, Pietrobon R, Hocker S, Guller U, Shankar A, Higgins LD. The relationship between surgeon and hospital volume and outcomes for shoulder arthroplasty. *J Bone Joint Surg*. 2004;86(3):496–505.
 57. Hervey S, Purves H, Guller U, Toth A, Vail T, Pietrobon R. Provider volume of total knee arthroplasties and patient outcomes in the HCUP-Nationwide Inpatient Sample. *J Bone Joint Surg Am Vol*. 2003;85(9):1775–83.

Users' Guide to the Surgical Literature: How to Assess a Pilot Trial in Surgery

12

Guowei Li, Xuerui Bai, Gillian A. Lancaster, and Lehana Thabane

Abbreviations

Aqueous-PREP	Preoperative Aqueous Anti-septic Skin Solution in Open Fractures
BMJ	<i>British Medical Journal</i>
BJC	<i>British Journal of Cancer</i>
BJOG	<i>British Journal of Obstetrics and Gynaecology</i>
BJS	<i>British Journal of Surgery</i>
CONSORT	Consolidated Standards of Reporting Trials
CV	cardiovascular
IDEAL	Idea, Development, Exploration, Assessment, and Long-term Follow-up
JAMA	<i>Journal of the American Medical Association</i>
NJEM	<i>New England Journal of Medicine</i>
NPWT	Negative Pressure Wound Therapy

RCT	Randomized Controlled Trial
TCM	Traditional Chinese Medicine
VFS	Videofluoroscopic Swallowing Study

Clinical Scenario

A hip fracture leads to significant pain, bleeding, immobility, and subsequent complications. Early surgery may reduce morbidity and mortality due to hip fractures. However, most patients with a hip fracture must wait a long time to receive surgery mainly because of delayed medical clearance and operation room access. Therefore, you hope to run a multicenter randomized clinical trial (RCT) to assess whether accelerated care (i.e., accelerated medical clearance plus surgical access) would lead to better outcomes compared with standard care (i.e., regular medical clearance plus surgery). But before you begin, you are uncertain about whether it is feasible to per-

G. Li
Center for Clinical Epidemiology and Methodology (CCEM), The Affiliated Guangdong Second Provincial General Hospital of Jinan University, Guangzhou, China
St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada
e-mail: ligw@gd2h.org.cn
X. Bai
Department of Epidemiology, School of Medicine, Jinan University, Guangzhou, China
e-mail: baixr@stu2022.jnu.edu.cn

G. A. Lancaster
School of Medicine, Keele University, Keele, UK
e-mail: g.lancaster@keele.ac.uk
L. Thabane (✉)
St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada
Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
e-mail: thabanl@mcmaster.ca

form such a large clinical trial. You are concerned that you may not be able to recruit enough willing participants and whether they will be able to complete the follow-up assessments. Therefore, you decide to review the literature to see if a pilot study that addresses your concerns has already been published from which you can glean some information about the feasibility of your potential trial.

We return to this scenario at the end of the chapter after addressing some general issues to consider when appraising or planning a pilot trial.

Definition of Pilot Trials

For a long time, there has been no consensus on the differences in definition of *pilot* and *feasibility* studies: some consider the terms pilot and feasibility as being synonymous, while others argue that these two terms have related but different meanings [1]. Multiple terms such as *pilot study*, *feasibility trial*, *pilot work*, *pilot trial*, and *feasibility investigation*, have been used without clear distinction between them [2]. However, in 2016, as part of the process of developing the CONSORT (Consolidated Standards of Reporting Trials) extension to pilot and feasibility trials [3, 4], Eldridge et al. introduced a conceptual framework, and published definitions of what constitutes a feasibility study and a pilot study. A feasibility study is viewed as an overarching concept of uncertainty about aspects of a future study within which three distinct types of study are identified: randomized pilot studies in which part or all of the aspects of the future study are carried out on a smaller scale to see if it can be done, non-randomized studies which are similar but exclude randomization of participants, and other types of feasibility studies where some element of the future study is being addressed (e.g., intervention development or acceptability) but no part of the future trial is being conducted [4, 5]. Within this framework, they define a feasibility study as a study asking “whether something can be done, should we proceed with it, and if so, how” and a pilot study as “a study in which a

future RCT or part of it, is conducted on a smaller scale” [4]. Essentially, Eldridge et al. appropriately note that “the corollary of these definitions is that all pilot studies (including trials) are feasibility studies but not all feasibility studies are pilot studies” [4].

In 2023, Bond et al. extended the conceptual framework of pilot and feasibility studies [6]. In their framework, the type of pilot trial was further discussed, where pilot trials can be categorized as either internal or external pilot trials. Data collected from an external pilot trial will not be used in the final analysis of any future main trial, while an internal pilot trial, designed as part of the future trial, will contribute its data to the final data analysis. In general, an internal pilot trial is performed when most feasibility uncertainties of conducting a main trial have been addressed, the main trial has started, and a pilot trial is needed to monitor and resolve any remaining uncertainties about recruitment and randomization. In this chapter we focus on external pilot and feasibility studies, and for brevity refer to any feasibility or pilot study as “pilot studies,” and any feasibility or pilot trial as “pilot trials” as appropriate.

Importance or Necessity of Conducting a Pilot Trial

Dependent upon the feasibility objective(s), reasons for conducting a pilot trial include procedural, resource- or management-related, scientific, and methodological rationales. In Thabane et al.’s paper, the importance of doing a pilot trial was well discussed [7]. In brief, a pilot study can help (1) assess whether the procedures (such as recruitment rates, adherence levels, and inclusion/exclusion criteria, among others) in a large-scale study are feasible; (2) evaluate and identify the issues related to resources and management including budgets, time/personnel, data collection/storage, and capacity in participating centers; (3) explore the scientific concerns such as preliminary safety evaluations, intervals/doses of interventions, and biological activity; and (4) investigate methodological issues, which includes assisting with sample size estimation for

large-scale studies, study design modifications, statistical plan adjustment, to mention a few.

For instance, the Aqueous-PREP (preoperative aqueous antiseptic skin solution in open fractures) pilot study was a cluster randomized crossover trial aiming to evaluate the feasibility of conducting a main trial to assess comparative effectiveness of a 10% povidone-iodine solution versus a 4% chlorhexidine gluconate solution for the management of open fractures [8]. There were five feasibility objectives including the implementation of the interventions during the run-in period, participant enrollment during the two 2-month recruitment phases, follow-up of participants, application of the trial interventions as per the cluster randomization crossover scheme, and the accurate documentation of the primary outcome of surgical site infection. It was found that all the five feasibility outcomes were satisfactorily met: all participants were correctly assigned during the run-in period, 92% of eligible participants were enrolled during recruitment phases, 98% of participants complied with the assigned treatment and completed follow-up, and the primary outcome was 100% accurately recorded. These results indicated the feasibility of a further large-scale trial to compare the effectiveness between the two types of solution for the management of open fractures.

Current Practice in Pilot Trials in the Literature

Given their notable importance and wide application in clinical areas, pilot studies have received little attention or suboptimal scrutiny in the scientific community. For example, publications of pilot studies rarely mention whether such studies are conducted to inform a subsequent large study, and journals adopt substantially varying editorial policies concerning publication of pilot studies. Lancaster et al. surveyed 7 journals (4 general medical journals: *BMJ* [British Medical Journal], *Lancet*, *JAMA* [Journal of the American Medical Association], and *NJEM* [New England Journal of Medicine]; and 3 subject-specific journals: *BJC* [British Journal of Cancer], *BJOG* [British

Journal of Obstetrics and Gynaecology], and *BJS* [British Journal of Surgery]) [9]. They found 90 pilot studies published in these journals for the years 2000 and 2001. However, only 4 (out of 90) pilot studies specifically stated that the pilot trial was in preparation for a main RCT. They also contacted the 7 journal editors and found that 4 (out of 7) journals had no publication policy for pilot studies, and 1 journal indicated the journal did not publish pilot studies. Arain et al. repeated this work to assess whether practice and editorial policy of pilot studies had changed over the years [10]. They reviewed the years 2007 and 2008, and found little change in practice compared with Lancaster et al.'s study [9]. Another editorial from the cardiovascular medicine journal *Circulation* reported that 41 pilot studies had been published since 2004 [11]. After peer-review, most of the authors were requested to explicitly define their trials as pilots to alert the audience to the preliminary nature of the findings including small sample size and lack of generalizability. This is not to be recommended when studies are designed as underpowered main studies looking at clinical outcomes of effectiveness, as the primary focus of pilot studies should be on feasibility and process outcomes. This is a common problem which has also been highlighted in the use of surrogate outcomes [12]. Li et al. conducted a review to assess the adherence to the guideline (CONSORT extension to pilot trials, published in 2016 [4]) in traditional Chinese medicine (TCM) by randomly selecting 50 TCM pilot trials (38 trials published before 2016, and 12 trials after 2016) [13]. They reported a median guideline adherence level of 68% (the first quartile–third quartile: 18–87%), with the lowest adherence of 4% found in item 6c (prespecified criteria used to judge progression to future definitive trial). After comparing the TCM pilot trials published before and after 2016, only the adherence level of item 12a (qualitative or quantitative methods used to address objectives) was significantly higher in trials published after 2016, indicating a minimal effect on guideline adherence exerted by the guideline. All these findings indicate the suboptimal nature of the current practices of pilot studies in the literature.

The IDEAL (idea, development, exploration, assessment, and long-term follow-up) framework was set up to provide guidance and recommendations for an integrated evaluation system for surgical studies [14]. The adoption of the IDEAL framework has been slow but is gaining increased momentum in the surgical literature [15]. The IDEAL framework is currently being updated [16], and is expected to explicitly incorporate pilot studies in the future under the “exploration” category. It is hoped that inclusion of pilot trials will address some of the inappropriate practices regarding the conduct and interpretation of small studies in surgery. For example, in surgical trials comparing two techniques, authors often conclude inappropriately that absence of evidence (i.e., p value >0.05 based on the data) is equivalent to evidence of absence (i.e., in fact no difference between the two techniques) [15]. This incorrect interpretation may mislead researchers to thinking that further properly designed studies are not needed to detect the differences between interventions.

Key Considerations for Designing a Pilot Trial

Before conducting a pilot study, some key elements should be carefully considered. Several key considerations we emphasize here include study rationale, objective(s), feasibility and patient-centered outcome(s), sample size justification, appropriate methods of analysis, and criteria of success.

Study Rationale

It is important to show readers why a pilot trial is needed before a future main RCT is conducted, given the current scientific background and available evidence. The rationale for a pilot study is generally to explore areas of uncertainty that need to be addressed before the large-scale trial can be planned [4], because according to the principles of the Helsinki declaration it is unethical to expose participants unnecessarily to the

unknown risks of research [17]. Based on the updated conceptual framework of pilot and feasibility studies [6], at early stages of developing interventions where maximum uncertainty exists, various methodological approaches could be used. For example, a pilot study can be conducted as a qualitative exploration of stakeholders’ views on acceptability of or specification for a proposed new intervention. When there is less uncertainty, a pilot study could be performed in which all or some parts of the future main trial are assessed for feasibility. When most uncertainty is resolved, a randomized external pilot trial is generally appropriate to include either all or a number of parts of the future main trial on a smaller scale, which if successful is usually the end of the feasibility pathway.

An example can be found in Dingemans et al.’s study that aimed to assess the feasibility of a new portable single-use negative pressure wound therapy (NPWT) device in patients undergoing major foot ankle surgery: “... *However, as only one prospective study on prophylactic negative pressure wound therapy in patients undergoing major lower extremity fracture surgery is available, evidence on its beneficial effect is limited and it is uncertain whether the extra expenses of the NPWT device are justified. Additionally, lately portable, single-use devices have been developed, which allow early discharge without the need for specialized home care. It is however unknown whether the good results observed with regular NPWT devices are equalled by these newly developed devices*” [18].

Objective(s) and Outcomes(s)

The primary objective(s) of a pilot study should concern the feasibility of performing a subsequent main study. To address the feasibility objective(s), prespecified outcome measurements should be carefully chosen. In general, outcomes of pilot studies may include recruitment rates, adherence levels, data completion, and variance estimates, among others. For instance, in Skoretz et al.’s study, the “... primary objective was to determine the feasibility of using validated and

objective interpretation measures for videofluoroscopy in conjunction with nasendoscopy to assess swallowing and upper airway physiology on prospectively enrolled CV [cardiovascular] surgery patients following prolonged intubation” and the “... secondary objective was to explore the tolerability and impact of this study on patients and nursing practice” [19]. Accordingly, their outcome measurements included “recruitment rate, patient participation, task completion durations, and the inter-rater reliability of VFS [videofluoroscopic swallowing study] measures using the intraclass correlation coefficient.”

Some pilot studies may be performed to determine an appropriate endpoint for the main study; therefore, researchers may not be able to choose the most appropriate primary outcome for the main study until the pilot trial is completed. Of note, however, researchers should explicitly state that identifying the primary outcome for the main study is one of the pilot study objectives. Likewise, exploring learning curve effects may also be an important objective in surgical pilot studies. Learning curves are a well-known way to increase surgical proficiency over time, but consequently these could distort comparisons between treatment groups [20, 21]. How to incorporate learning curve effects in surgical trials is an important issue that requires further investigation.

Sample Size

One key requirement of pilot studies is that they should ensure a sufficient sample of patients large enough to provide adequate information about the feasibility of conducting a large-scale RCT. For pilot trials, sample size calculations should always be based on the feasibility objectives and scientific rationale. In many cases, pilot studies give a justified rationale as to the number in the sample without performing a formal sample size calculation. For example, when testing the understanding and acceptability of a questionnaire 10 patients may be deemed adequate. In other studies that have a primary objective of achieving an estimated recruitment,

retention, or adherence rate, a desired degree of precision (variation, or confidence interval) around the estimated rate is usually needed to help calculate the sample size [22]. We would refer readers to Thabane et al.'s publication about how to use the confidence interval approach to estimate sample size for a pilot study where rates are to be estimated [7].

Pilot studies may also provide some information about estimating variation of an effect size to inform the sample size calculation for a future main RCT. Readers can refer to relevant methodological publications for more detail [23, 24]. Nevertheless, caution is needed because findings from pilot studies may easily mislead sample size calculations for a main trial if the estimates are based on a small sample [25].

There are some recent methodological papers discussing the sample size calculation for process outcomes based on progression criteria and for external pilot cluster trials with binary feasibility outcomes [26, 27]. Other considerations for sample size calculation for pilot studies in the literature include (1) discussion with health professionals for further supplemental information, for example, on availability of patients; (2) creating a sample size table with a range of adjusted variation estimates of an effect size to emphasize how the numbers change with differing levels of uncertainty around the outcome of interest; and (3) running simulation studies to assess different scenarios and present a range of results with estimates of variability [7, 24, 28].

Methods of Analysis

Pilot studies should not primarily focus on treatment effects and tests of effectiveness; instead, they are conducted to address the feasibility of conducting a main RCT. Therefore, statistical tests of significance may not be needed or necessary. Where hypothesis tests are performed a sample size calculation should have been carried out or some justification given. It is also helpful to include a cautionary caveat about the study being underpowered and therefore the results should be treated as preliminary [9]. Findings

from pilot studies are usually summarized descriptively by such as means with standard deviations and counts with percentages for each treatment group separately. Moreover, it is not uncommon to show results of pilot studies qualitatively; for example, by using narrative descriptions. An example can be found in Forero et al.'s study, in which "the feasibility outcomes were reported descriptively and narratively. For the clinical endpoints, only descriptive statistics, mean (standard deviation) for continuous outcomes and raw count (%) for categorical outcomes, were reported. Due to the nature of pilot designs, we chose not to conduct any informative statistical tests on the collected data" [29].

Criteria of Success

As mentioned above, pilot studies are not conducted to estimate an effect size or compare different treatment effects due to their small sample size and lack of generalizability. However, it is good practice to draw up progression criteria before the pilot study begins that relate to objectives such as recruitment rates or adherence rates. As a rule of thumb, the criteria of success should be related to the primary feasibility objective(s). Therefore, feasibility findings from pilot trials can inform progression to a subsequent main RCT including whether a large RCT should be planned as is (main trial feasible, no modification needed), planned but with changes (main trial feasible, protocol amendments needed), or not planned (main trial not feasible, aborted completely) [26]. As mentioned above, aspects of the feasibility may include procedural, resource- or management-related, scientific, and methodological considerations. However, two recent studies reported that progression criteria of success of pilot trials were not adequately prespecified in many pilot trial reports or protocols, indicating room for improvement of progression criteria in pilot trials [30, 31].

Recently, Mellor et al. provided several recommendations for progression criteria during external randomized pilot trial design, conduct, analysis, and reporting [32]. During the design

stage of a pilot trial, they recommend that researchers "consider progression criteria from the earliest opportunity; map progression criteria to feasibility objectives; consider quantitative and qualitative interpretations of feasibility; provide justification; develop guidelines rather than rules; and seek input from relevant stakeholders." During pilot trial conduct, they recommend that pilot trial data are regularly monitored against the progression criteria. At the stage of analysis, researchers are recommended to "avoid considering each progression criterion in isolation; engage in discussion with relevant stakeholders; consider context and other factors external to the pilot trial; and consider feasibility (can we?) and progression (will we?)." When reporting the pilot trial, it is recommended to provide relevant information on progression criteria in a table format in the methods (prespecified progression criteria, rationale for such criteria, brief description of the involvement of stakeholders in developing or agreeing these criteria, and proposed method of multi-criteria assessment) and in the results (any changes with reasons to progression criteria after the trial started, findings of progression criteria related to feasibility, implications of these findings, and overall statement of trial feasibility).

Reporting of a Pilot Trial

Regarding the reporting of pilot studies, we would refer readers to the published CONSORT extension for pilot and feasibility trials [4]. The guideline provides instructions on how to transparently and adequately report abstracts and main texts of pilot studies with detailed examples and explanations. It also compares the standard CONSORT checklist items with the extension checklist items for pilot studies, with some items not applicable to pilot studies and some new items added.

The CONSORT extension for pilot and feasibility trials mainly focuses on randomized pilot studies. Therefore, in 2019, two of the authors provided some general guidance on how to report non-randomized pilot studies, where the most common types of non-randomized pilot studies

included intervention development, patient-reported outcome measures development, piloting some components of the main trial, implementation of research findings, feasibility studies in preparation for a cohort or other large-scale study, and feasibility studies that test preliminary hypotheses of associations [33].

The CONSORT extension also covers the registration and protocol of a pilot trial (items no. 23 and 24) [4]. Although there is no specific registry or platform for pilot trials, trial authors are encouraged to register their pilot trials in a similar way to other types of trials before a pilot trial commences. The authors also provided general guidance on how to report protocols of pilot trials, in which the guidance incorporated the CONSORT extension for pilot and feasibility trials and the SPIRIT guideline (Standard Protocol Items: Recommendations for Interventional Trials) for main trials [34].

We suggest researchers who plan to design, conduct, and report a pilot study should carefully refer to the reporting guidelines and adhere to the checklist items as appropriate. It is important to note that the CONSORT extension checklist items and the guidance on non-randomized pilot trials and protocols of pilot trials are very useful reference guides to consult at the design stage of the pilot study and many examples are given in these papers. They also serve as robust appraisal tools for assessing published pilot studies. In addition, many examples across all disciplines are found in the *Pilot and Feasibility Studies* journal, which includes three special series on Intervention Development, piloting Patient-Reported Outcome Measures, and testing the feasibility of Implementation Science initiatives.

Key Resources for Further References

To assist with the implementation of pilot studies in surgery, Table 12.1 summarizes some key resources for readers. The table covers the aspects of surgical pilot studies including their reporting guidelines, definition and framework, general conduct (the what/why/ how and common misconceptions), deficiencies in current practice, the IDEAL framework and recommendations, ethical and sample size issues, progression criteria in pilot trials, more examples of randomized and non-randomized surgical pilot trials, and other helpful resources. For instance, Thabane et al.'s publication [7] gives a straightforward tutorial on the what/why/how regarding a pilot trial; it also covers some aspects including the differences between pilot trials and proof-of-concept studies, some common confusions and misconceptions, and some frequently asked questions. Ethical issues of pilot studies have not been appropriately addressed in the literature. We did a quick literature search of ethics guidelines for research, but did not find any detailed information on (1) whether a pilot study is ethical if its feasibility cannot be guaranteed or (2) how to approach participants for their consent given the feasibility nature of pilot trials. However, we would refer readers to Thabane et al.'s publication [7] and the CONSORT reporting guideline by Eldridge et al. [4] that proposed some solutions to such ethical issues. Examples of surgical pilot trials including randomized [8, 35] and non-randomized pilot trials [36, 37], as well as examples of published protocols [38–42], can also be found in Table 12.1.

Table 12.1 Some key resources for readers’ further references to surgical pilot studies

Aspects of pilot studies	Key reference ^a
<i>Reporting guidelines</i>	
CONSORT extension to pilot trials	Eldridge, 2016 [4]
How to report non-randomized pilot and feasibility studies	Lancaster, 2019 [33] ^b
Pilot and feasibility study protocols	Thabane, 2019 [34] ^b
<i>Definition and framework</i>	
<i>The what/why/how and common misconceptions</i>	
<i>Deficiencies in current practice</i>	
<i>The IDEAL framework and recommendations</i>	
<i>Ethical issues</i>	
<i>Sample size issues</i>	
<i>Progression criteria in pilot trials</i>	
<i>Examples of randomized surgical pilot trials</i>	
Surgical pilot trial reports	Horth, 2021 [35] ^b ; Sprague, 2021 [8] ^b
Protocol of surgical pilot trials	Snee, 2016 [38]; Kearney, 2017 [39]; Kirkegaard-Klitbo, 2023 [40] ^b
<i>Examples of non-randomized surgical pilot trials</i>	
Surgical pilot trial reports	Mason, 2015 [36]; Lim, 2018 [37] ^b
Protocol of surgical pilot trials	Chang, 2021[41] ^b ; Moroi, 2022 [42] ^b
<i>Other helpful resources</i>	
Pilot and Feasibility Studies Journal	https://pilotfeasibilitystudies.biomedcentral.com/
Resources on pilot and feasibility studies website	Home—Pilot and feasibility studies (qmul.ac.uk)
Special series as general examples	https://www.biomedcentral.com/collections/interventiondevelopment https://www.biomedcentral.com/collections/implementation-science-pilot-studies https://www.biomedcentral.com/collections/pilotfeasibilityPROMs

^aShown as: first author, publication year [study reference]
^bUpdated references from the first edition

Returning to the Scenario

After reviewing the literature (Appendix Table shows the detailed search strategy in the PubMed), you identified a pilot trial (named *the HIP ATTACK pilot trial*) comparing accelerated care with standard care in patients with a hip fracture [45]. Based on the CONSORT extension guideline to pilot trials, you appraise the article carefully.

In general, the article was adequately conducted and well reported (Table 12.2 shows the items reported in the article according to the CONSORT checklist), supporting the

validity of their study findings. All the feasibility objectives were met in the pilot trial; that is, 60 patients were recruited, 80% of eligible patients underwent randomization, and 100% of patients completed the follow-up assessments. The pilot trial was carried out in a very similar setting to your own and suggested that a large-scale trial would be feasible. You discover from colleagues that the authors have in fact gone ahead with a large-scale definitive trial to answer the clinical question, which is currently recruiting. You decide to contact the team to see if you can become involved in the study.

Table 12.2 CONSORT checklist of information reported in the HIP ATTACK pilot trial^a

Section/topic	Item no.	Checklist item	Reported (Yes/No/NA)
<i>Title and abstract</i>			
	1a	Identification as a pilot or feasibility randomized trial in the title	Yes
	1b	Structured summary of pilot trial design, methods, results, and conclusions (for specific guidance, see CONSORT abstract extension for pilot trials)	Yes
<i>Introduction</i>			
Background and objectives	2a	Scientific background and explanation of rationale for future definitive trial, and reasons for randomized pilot trial	Yes
	2b	Specific objectives or research questions for pilot trial	Yes
<i>Methods</i>			
Trial design	3a	Description of pilot trial design (such as parallel, factorial) including allocation ratio	Yes
	3b	Important changes to methods after pilot trial commencement (such as eligibility criteria), with reasons	Yes
Participants	4a	Eligibility criteria for participants	Yes
	4b	Settings and locations where the data were collected	Yes
	4c	How participants were identified and consented	Yes
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	Yes
Outcomes	6a	Completely defined prespecified assessments or measurements to address each pilot trial objective specified in 2b, including how and when they were assessed	Yes
	6b	Any changes to pilot trial assessments or measurements after the pilot trial commenced, with reasons	NA
	6c	If applicable, prespecified criteria used to judge whether, or how, to proceed with future definitive trial	NA
Sample size	7a	Rationale for numbers in the pilot trial	Yes
	7b	When applicable, explanation of any interim analyses and stopping guidelines	NA
<i>Randomization</i>			
Sequence generation	8a	Method used to generate the random allocation sequence	Yes
	8b	Type of randomization(s); details of any restriction (such as blocking and block size)	Yes
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	Yes
Implementation	10	Who generated the random allocation sequence, who enrolled participants and who assigned participants to interventions	Yes
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how	Yes
	11b	If relevant, description of the similarity of interventions	Yes
Statistical methods	12	Methods used to address each pilot trial objective whether qualitative or quantitative	Yes

(continued)

Table 12.2 (continued)

Section/topic	Item no.	Checklist item	Reported (Yes/No/NA)
<i>Results</i>			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were approached or assessed for eligibility, randomly assigned, received intended treatment, and were assessed for each objective	Yes
	13b	For each group, losses and exclusions after randomization, together with reasons	Yes
Recruitment	14a	Dates defining the periods of recruitment and follow-up	Yes
	14b	Why the pilot trial ended or was stopped	Yes
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	Yes
Numbers analyzed	16	For each objective, number of participants (denominator) included in each analysis. If relevant, these numbers should be by randomized group	Yes
Outcomes and estimation	17	For each objective, results including expressions of uncertainty (such as 95% confidence interval) for any estimates. If relevant, these results should be by randomized group	Yes
Ancillary analyses	18	Results of any other analyses performed that could be used to inform the future definitive trial	Yes
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	Yes
	19a	If relevant, other important unintended consequences	NA
<i>Discussion</i>			
Limitations	20	Pilot trial limitations, addressing sources of potential bias and remaining uncertainty about feasibility	Yes
Generalizability	21	Generalizability (applicability) of pilot trial methods and findings to future definitive trial and other studies	Yes
Interpretation	22	Interpretation consistent with pilot trial objectives and findings, balancing potential benefits and harms, and considering other relevant evidence	Yes
	22a	Implications for progression from pilot to future definitive trial, including any proposed amendments	Yes
<i>Other information</i>			
Registration	23	Registration number for pilot trial and name of trial registry	Yes
Protocol	24	Where the pilot trial protocol can be accessed, if available	NA
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	Yes
	26	Ethical approval or approval by research review committee, confirmed with reference number	Yes

Eldridge et al. [4]

*We strongly recommend reading this statement in conjunction with the CONSORT 2010, extension to randomized pilot and feasibility trials, explanation and elaboration for important clarifications on all the items. If relevant, we also recommend reading CONSORT extensions for cluster randomized trials, non-inferiority and equivalence trials, non-pharmacological treatments, herbal interventions, and pragmatic trials. Additional extensions are forthcoming: for those and for up-to-date references relevant to this checklist, see www.consort-statement.org

Concluding Remarks

Pilot studies provide a platform to evaluate feasibility and explore uncertainties in preparation for a subsequent large-scale RCT. In most

cases, it is very important and necessary to perform a pilot study before running a main trial. However, some key elements have to be carefully considered when designing a pilot study as the objectives are different to those of the main study.

Moreover, researchers who plan to design, implement, and report a pilot study are strongly advised to refer to the CONSORT extension reporting guideline and adhere to the statement and checklist items as appropriate.

Appendix Table: Detailed Literature Search Strategy Used in the PubMed from January 1, 2000 to March 10, 2024

Step	Detailed search strategies	No. of records
#1	"Hip Fractures"[Mesh]	20,614
#2	"Hip Fractures"[Title/Abstract]	9235
#3	"Trochanteric Fractures"[Title/Abstract]	763
#4	"Subtrochanteric Fracture*" [Title/Abstract]	794
#5	"Femoral Fractures"[Mesh]	29,502
#6	"Femoral Fracture*" [Title/Abstract]	6544
#7	"Standard care*" [Title/Abstract]	13,853
#8	"Accelerated care*" [Title/Abstract]	31
#9	"Pilot Projects"[Mesh]	127,864
#10	"Feasibility Studies"[Mesh]	78,424
#11	"Pilot Project*" [Title/Abstract]	5089
#12	"Pilot Stud*" [Title/Abstract]	111,897
#13	"Pilot Trial*" [Title/Abstract]	7098
#14	"Feasibility Stud*" [Title/Abstract]	21,914
#15	"Feasibility Trial*" [Title/Abstract]	2717
#16	#1 OR #2 OR #3 OR #4 OR #5 OR #6	34,921
#17	#9 OR #10	197,418
#18	#11 OR #12 OR #13 OR #14 OR #15	143,846
#19	#17 OR #18	247,287
#20	#7 AND #8	6
#21	#16 AND #19 AND #20	1

References

- Whitehead AL, Sully BG, Campbell MJ. Pilot and feasibility studies: is there a difference from each other and from a randomised controlled trial? *Contemp Clin Trials*. 2014;38(1):130–3.
- Arnold DM, Burns KE, Adhikari NK, Kho ME, Meade MO, Cook DJ. The design and interpretation of pilot trials in clinical research in critical care. *Crit Care Med*. 2009;37(1 Suppl):S69–74.

- Thabane L, Hopewell S, Lancaster GA, Bond CM, Coleman CL, Campbell MJ, Eldridge SM. Methods and processes for development of a CONSORT extension for reporting pilot randomized controlled trials. *Pilot Feasibility Stud*. 2016;2:25.
- Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, Lancaster GA. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ (Clin Res Ed)*. 2016;355:i5239.
- Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, Bond CM. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. *PLoS One*. 2016;11(3):e0150205.
- Bond C, Lancaster GA, Campbell M, Chan C, Eddy S, Hopewell S, Mellor K, Thabane L, Eldridge S. Pilot and feasibility studies: extending the conceptual framework. *Pilot Feasibility Stud*. 2023;9(1):24.
- Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, Robson R, Thabane M, Giangregorio L, Goldsmith CH. A tutorial on pilot studies: the what, why and how. *BMC Med Res Methodol*. 2010;10:1.
- Sprague S, Guyatt P, Bzovsky S, Nguyen U, Bhandari M, Thabane L, Petrisor B, Johal HS, Leonard J, Dodds S, Mossuto F, O'Toole RV, Howe A, Demyanovich HK, Camara M, O'Hara NN, Slobogean GP, The PREP-IT Investigators. Pragmatic randomized trial evaluating pre-operative aqueous antiseptic skin solution in open fractures (Aqueous-PREP): the feasibility of a cluster randomized crossover study. *Pilot Feasibility Stud*. 2021;7(1):61.
- Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract*. 2004;10(2):307–12.
- Araim M, Campbell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Med Res Methodol*. 2010;10(1):67.
- Loscalzo J. Pilot trials in clinical research: of what value are they? *Circulation*. 2009;119(13):1694–6.
- Campbell MJ, Lancaster GA, Eldridge SM. A randomised controlled trial is not a pilot trial simply because it uses a surrogate endpoint. *Pilot Feasibility Stud*. 2018;4:130.
- Li G, Wu D, Chen X, Zeng J, Li Z, Thabane L. Pilot and feasibility trials in traditional Chinese medicine: a literature review of current practice. *Pilot Feasibility Stud*. 2020;6:48.
- McCulloch P, Altman DG, Campbell WB, Flum DR, Glasziou P, Marshall JC, Nicholl J, Aronson JK, Barkun JS, Blazeby JM, Boutron IC, Campbell WB, Clavien PA, Cook JA, Ergina PL, Feldman LS, Flum DR, Maddern GJ, Nicholl J, Reeves BC, Seiler CM, Strasberg SM, Meakins JL, Ashby D, Black N, Bunker J, Burton M, Campbell M, Chalkidou K, Chalmers I, de Leval M, Deeks J, Ergina PL, Grant A, Gray M, Greenhalgh R, Jenicek M, Kehoe S, Lilford R, Littlejohns P, Loke Y, Madhock R, McPherson K, Meakins J, Rothwell P, Summerskill B, Taggart D, Tekkis P, Thompson M, Treasure T, Trohler U,

- Vandenbroucke J. No surgical innovation without evaluation: the IDEAL recommendations. *Lancet* (London, England). 2009;374(9695):1105–12.
15. McCulloch P, Feinberg J, Philippou Y, Kolias A, Kehoe S, Lancaster G, Donovan J, Petrinic T, Agha R, Pennell C. Progress in clinical research in surgery and IDEAL. *Lancet* (London, England). 2018;392:88.
 16. Hirst A, Philippou Y, Blazeby J, Campbell B, Campbell M, Feinberg J, Rovers M, Blencowe N, Pennell C, Quinn T, Rogers W, Cook J, Kolias AG, Agha R, Dahm P, Sedrakyan A, McCulloch P. No surgical innovation without evaluation: evolution and further development of the IDEAL framework and recommendations. *Ann Surg*. 2019;269(2):211–20.
 17. World Medical Association declaration of Helsinki. Recommendations guiding physicians in biomedical research involving human subjects. *JAMA*. 1997;277(11):925–6.
 18. Dingemans SA, Birnie MF, Backes M, de Jong VM, Luitse JS, Goslings JC, Schepers T. Prophylactic negative pressure wound therapy after lower extremity fracture surgery: a pilot study. *Int Orthop*. 2018;42:1–7.
 19. Skoretz SA, Yau TM, Granton JT, Martino R. The feasibility of assessing swallowing physiology following prolonged intubation after cardiovascular surgery. *Pilot Feasibility Stud*. 2017;3:62.
 20. Cook JA, Ramsay CR, Fayers P. Statistical evaluation of learning curve effects in surgical trials. *Clin Trials* (London, England). 2004;1(5):421–7.
 21. Simpson AH, Howie CR, Norrie J. Surgical trial design – learning curve and surgeon volume: determining whether inferior results are due to the procedure itself, or delivery of the procedure by the surgeon. *Bone Joint Res*. 2017;6(4):194–5.
 22. Cocks K, Torgerson DJ. Sample size calculations for pilot randomized trials: a confidence interval approach. *J Clin Epidemiol*. 2013;66(2):197–201.
 23. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. *Pharm Stat*. 2005;4(4):287–91.
 24. Teare MD, Dimairo M, Shephard N, Hayman A, Whitehead A, Walters SJ. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials*. 2014;15:264.
 25. Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch Gen Psychiatry*. 2006;63(5):484–9.
 26. Lewis M, Bromley K, Sutton CJ, McCray G, Myers HL, Lancaster GA. Determining sample size for progression criteria for pragmatic pilot RCTs: the hypothesis test strikes back! *Pilot Feasibility Stud*. 2021;7(1):40.
 27. Hemming K, Taljaard M, Gkini E, Bishop J. Sample size determination for external pilot cluster randomised trials with binary feasibility outcomes: a tutorial. *Pilot Feasibility Stud*. 2023;9(1):163.
 28. Sim J, Lewis M. The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency. *J Clin Epidemiol*. 2012;65(3):301–8.
 29. Forero M, Heikkilä A, Paul JE, Cheng J, Thabane L. Lumbar transversus abdominis plane block: the role of local anesthetic volume and concentration—a pilot, prospective, randomized, controlled trial. *Pilot Feasibility Stud*. 2015;1:10.
 30. Mbuagbaw L, Kosa SD, Lawson DO, Stalteri R, Olaiya OR, Alotaibi A, Thabane L. The reporting of progression criteria in protocols of pilot trials designed to assess the feasibility of main trials is insufficient: a meta-epidemiological study. *Pilot Feasibility Stud*. 2019;5:120.
 31. Mellor K, Eddy S, Peckham N, Bond CM, Campbell MJ, Lancaster GA, Thabane L, Eldridge SM, Dutton SJ, Hopewell S. Progression from external pilot to definitive randomised controlled trial: a methodological review of progression criteria reporting. *BMJ Open*. 2021;11(6):e048178.
 32. Mellor K, Albury C, Dutton SJ, Eldridge S, Hopewell S. Recommendations for progression criteria during external randomised pilot trial design, conduct, analysis and reporting. *Pilot Feasibility Stud*. 2023;9(1):59.
 33. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. *Pilot Feasibility Stud*. 2019;5:114.
 34. Thabane L, Lancaster G. A guide to the reporting of protocols of pilot and feasibility trials. *Pilot Feasibility Stud*. 2019;5:37.
 35. Horth D, Sanh W, Moisiuk P, O'Hare T, Shargall Y, Finley C, Hanna W, Agzarian J, Forero M, Davis K, Vanniyasingam T, Thabane L, Shanthanna H. Continuous erector spinae plane block versus intercostal nerve block in patients undergoing videoassisted thoracoscopic surgery: a pilot randomized controlled trial. *Pilot Feasibility Stud*. 2021;7(1):56.
 36. Mason JD, Blencowe NS, McNair AG, Stevens DJ, Avery KN, Pullyblank AM, Blazeby JM. Investigating the collection and assessment of patient-reported outcome data amongst unplanned surgical hospital admissions: a feasibility study. *Pilot Feasibility Stud*. 2015;1:16.
 37. Lim CP, Roberts M, Chalhoub T, Waugh J, Delegate L. Cadaveric surgery in core gynaecology training: a feasibility study. *Gynecol Surg*. 2018;15(1):4.
 38. Snee M, McParland L, Collinson F, Lowe C, Striha A, Baldwin D, Naidu B, Sebag-Montefiore D, Gregory W, Bestall J. The SABRTooth feasibility trial protocol: a study to determine the feasibility and acceptability of conducting a phase III randomised controlled trial comparing stereotactic ablative radiotherapy (SABR) with surgery in patients with peripheral stage I non-small cell lung cancer (NSCLC) considered to be at higher risk of complications from surgical resection. *Pilot Feasibility Stud*. 2016;2(1):5.

39. Kearney RS, Parsons N, Mistry D, Young J, Brown J, O'Beirne-Elliman J, Costa M. A protocol for a feasibility randomised controlled trial to assess the difference between functional bracing and plaster cast for the treatment of ankle fractures. *Pilot Feasibility Stud.* 2017;3(1):11.
40. Kirkegaard-Klitbo A, Shabanzadeh DM, Olsen MH, Lindschou J, Gluud C, Sørensen LT. One-step laparoscopic cholecystectomy with common bile duct exploration and stone extraction versus two-step endoscopic retrograde cholangiography with stone extraction plus laparoscopic cholecystectomy for patients with common bile duct stones: a randomised feasibility and pilot clinical trial-the preGallStep trial. *Pilot Feasibility Stud.* 2023;9(1):21.
41. Chang SJ, Cajigas I, Guest JD, Noga BR, Widerström-Noga E, Haq I, Fisher L, Luca CC, Jagid JR. Deep brain stimulation of the Cuneiform nucleus for levodopa-resistant freezing of gait in Parkinson's disease: study protocol for a prospective, pilot trial. *Pilot Feasibility Stud.* 2021;7(1):117.
42. Moroi R, Shiga H, Nochioka K, Shimoyama Y, Kuroha M, Kakuta Y, Kinouchi Y, Masamune A. A pilot study investigating the safety and feasibility of endoscopic dilation using a radial incision and cutting technique for benign strictures of the small intestine: a study protocol. *Pilot Feasibility Stud.* 2022;8(1):85.
43. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for designing and evaluating feasibility pilot studies. *Med Care.* 2022;60(1):95–103.
44. Khan MIU, Mbuagbaw L, Holec M, Bdair F, Durrani ZH, Mellor K, Eddy S, Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Lancaster GA, Thabane L. Transparency of informed consent in pilot and feasibility studies is inadequate: a single-center quality assurance study. *Pilot Feasibility Stud.* 2021;7(1):96.
45. Accelerated care versus standard care among patients with hip fracture: the HIP ATTACK pilot trial. *CMAJ.* 2014;186(1):E52–60.

Assessing Noninferiority Randomized Controlled Surgical Trials

13

Alla Bengizi, Mansi Patel, Sandra Ofori,
Yaadwinder Shergill, and Jason W. Busse

Abbreviations

APM	Arthroscopic Partial Menis- ectomy
CI	Confidence Interval
ESCAPE	Effect of Early Surgery vs Physical Therapy on Knee Function Among Patients with Nonobstructive Meniscal Tears
IKDC	International Knee Documen- tation Committee
IMPACT	Initiative on Methods, Mea- surement, and Pain Assess- ment in Clinical Trials
ITT	Intention-to-Treat
PICO	Population, Intervention, Comparison, Outcome
PT	Physical Therapy
RCT	Randomized Controlled Trial

Clinical Scenario

During the weekly academic rounds at your institution, a discussion unfolds between a resident and a senior orthopedic surgeon regarding the comparative effectiveness of arthroscopic partial meniscectomy (APM) versus physical therapy (PT) for managing meniscal tears. They each defend an approach, and the issue of which is preferable is not resolved. The following day, while rounding in the clinic, you encounter a 58-year-old woman with a nonobstructive meniscal tear. Sensing her apprehension about potential surgery, you explain the various treatment options, including surgical and nonsurgical options. You schedule a follow-up appointment to reconvene with her in 1 week to decide on the approach she would prefer, and while awaiting the appointment you delve into the literature to prepare for this discussion.

A. Bengizi
Department of Health Research Methods,
Evidence, and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: bengizia@mcmaster.ca

M. Patel
Department of Surgery, McMaster University,
Hamilton, ON, Canada
e-mail: patelm96@mcmaster.ca

S. Ofori
Department of Medicine, McMaster University,
Hamilton, ON, Canada
e-mail: oforis@mcmaster.ca

Y. Shergill
Ottawa Hospital Research Institute,
Ottawa, ON, Canada

J. W. Busse (✉)
Department of Health Research Methods,
Evidence, and Impact, McMaster University,
Hamilton, ON, Canada

Department of Anesthesia, McMaster University,
Hamilton, ON, Canada
e-mail: bussejw@mcmaster.ca

Finding the Evidence

To better understand the current evidence for APM or PT for nonobstructive meniscus tear, you conduct a literature search in accordance with the steps outlined in Chap. 4. You structure your question according to the population, intervention, control, and outcome (PICO) format to identify key words in the search process. Your population of interest is adult patients with non-obstructive meniscal tears. The intervention of interest is PT, as compared to the current standard of care (APM). Your patient-important outcome of interest is knee function.

Using the search terms “nonobstructive meniscal tears” AND “arthroscopic partial meniscectomy” AND “physical therapy,” you search PubMed (<https://pubmed.ncbi.nlm.nih.gov/>) for relevant studies (Appendix A). This search returns four articles, one of which is a randomized clinical trial titled “Effect of Early Surgery vs Physical Therapy on Knee Function Among Patients with Nonobstructive Meniscal Tears (ESCAPE),” which addresses your clinical questions, and its noninferiority design caught your interest [1].

Randomized Controlled Trials

Randomized controlled trials (RCTs) are considered the gold standard for evaluating the effectiveness and safety of novel interventions [2]. Random allocation of participants to competing interventions balances prognostic factors and provides confidence that differences in outcomes can be attributed to the treatment under investigation [3]. RCTs can be designed to explore superiority, equivalence, or noninferiority, depending on the context and purpose of the study. For more information on RCTs, please refer to Chaps. 11, 12, 14, and 15.

Noninferiority RCTs explore whether a new therapy is not appreciably worse than a comparator, typically the current standard of care. The null hypothesis is that the new intervention is worse than standard care by greater than an acceptable amount, defined as the noninferiority

margin (Δ) [4]. A noninferiority design is appropriate when a new procedure is associated with fewer harms, costs, or patient burden [4, 5].

The noninferiority margin is integral to ensuring that the comparison between the new treatment and the existing one is valid and clinically meaningful. The United States Food and Drug Administration has suggested use of the fixed-margin approach, which identifies the smallest plausible benefit of the current standard of care [6]. The smallest plausible benefit can be estimated by conducting a meta-analysis of trials reporting the treatment effect of the intervention in question, which will be associated with a measure of precision—the 95% confidence interval (95% CI). The lower limit of the 95% CI can be used to estimate the smallest plausible benefit [6, 7]. Clinical judgment is then used to determine the proportion of the smallest plausible benefit that must be retained to conclude noninferiority. Key factors to consider when determining the preserved fraction include outcome severity, intervention efficacy, risk-benefit, and cost [7]. Commonly, 50% preserved fraction of the smallest plausible benefit should be retained. The resulting fraction of the smallest plausible benefit would then be concluded as the noninferiority margin [6, 7].

To illustrate an example of the fixed-margin approach, consider the effect of percutaneous coronary angioplasty versus coronary artery bypass grafting on all-cause mortality in patients with left main coronary artery disease. If the current literature shows that the standard surgical procedure—in this case the coronary artery bypass grafting, provides an absolute reduction of 12% (95% CI of 8–16%) in all-cause mortality compared to nonsurgical care, the smallest plausible benefit would be equal to 8%. Using the common preserved fraction of 50%, the noninferiority margin would be 4%. With this noninferiority margin, the null hypothesis would be rejected when the associated measure of precision does not exceed 4% (e.g., a point estimate demonstrating a 2% in the reduction in mortality in participants receiving coronary artery bypass grafting compared to percutaneous coronary angioplasty, with a 95% CI of 1–3%). Estimates

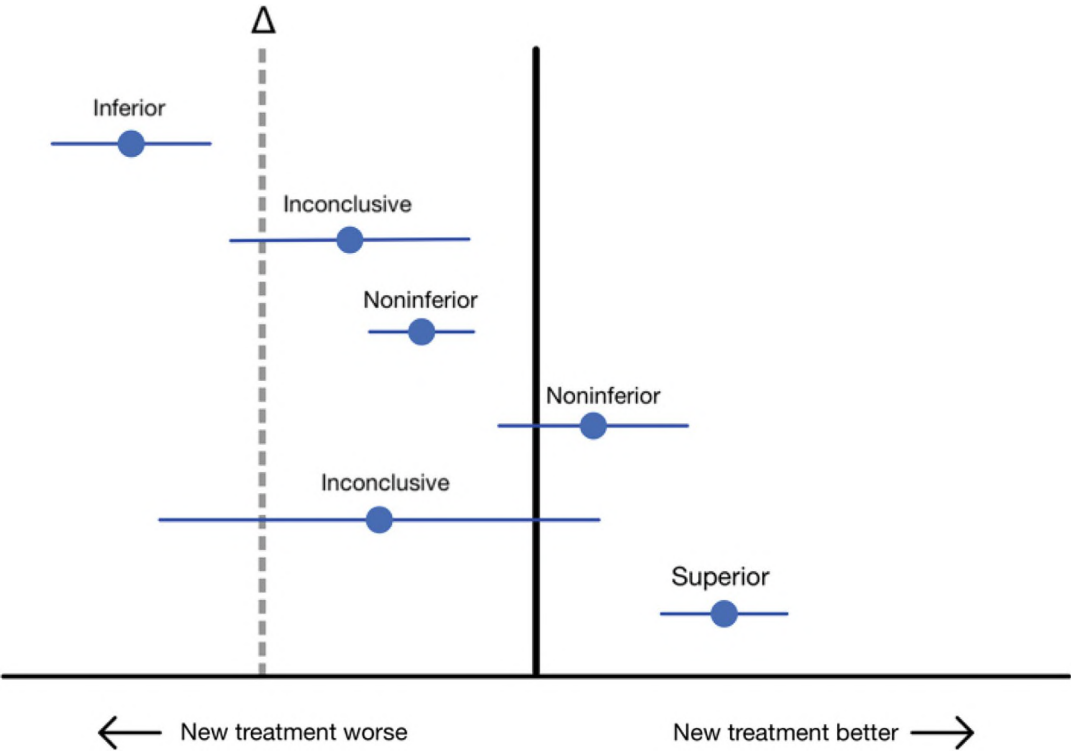


Fig. 13.1 Potential outcomes for the detected mean difference between two interventions. The mean is symbolized by the circle, while the bar illustrates the two-sided 95% confidence interval. The solid vertical line signifies

no difference between the two interventions, while the dashed line represents the noninferiority threshold (Δ). (This image has been adapted from Piaggio [8])

of precision that include the noninferiority margin are inconclusive, whereas those that include only benefit versus the comparator include superiority (Fig. 13.1).

If noninferiority trials choose overly generous margins (e.g., willing to cede considerable benefit), they may conclude noninferiority when many informed patients would be unwilling to pursue the new procedure given the largest possible decreased effectiveness associated with its use. Unfortunately, many trialists do not adequately report and justify their choice of noninferiority margin [9]. As a result, consumers of noninferiority trials should use their judgment as to whether non-inferiority has been established, rather than relying solely on the conclusions of the study authors.

You determine that the ESCAPE trial may be helpful to inform the upcoming discussion with your patient, and you undertake critical appraisal

of the article using a framework set forth by previous users' guide articles (Box 13.1) [10].

Box 13.1: Critical Appraisal Framework for Non-inferiority Randomized Controlled Trials

STEP 1: Are the results valid?

- Were the novel and standard surgical intervention groups prognostically similar at baseline? Was prognostic balance maintained as the trial progressed?
- Did the investigators analyze participants according to the surgical treatment they received, as well as the groups to which they were assigned?
- Did the investigators report a pre-defined noninferiority margin and

(continued)

Box 13.1 (continued)

guard against an unwarranted conclusion of noninferiority?

- Did the investigators power the study for a test of noninferiority?

STEP 2: What are the results?

- Were all patient-important outcomes considered?
- Were the results precise?
- Were the investigators appropriately interpreting the concept of noninferiority?

STEP 3: How can I apply the results of my patient or clinical practice?

- Were the participants similar to my patient?
- Are the likely advantages of the novel surgical treatment worth the potential harm and costs?

STEP 1: Are the Results Valid?

The ESCAPE trial was a noninferiority, multicenter, randomized clinical trial conducted in nine hospitals in the Netherlands, which included 321 participants aged 45–70 years with nonobstructive meniscal tears (i.e., no locking of the knee joint) [1]. Participants were randomized to either APM ($n = 159$) or PT ($n = 162$). The flow of participants in the ESCAPE trial is summarized in Fig. 13.2 [1].

Were the Novel and Standard Surgical Intervention Groups Prognostically Similar at Baseline? Was Prognostic Balance Maintained as the Trial Progressed?

Evaluating the similarity of prognostic factors between the two intervention groups is crucial to minimize the risk of selection, allocation, and confounding biases. Additional measures to reduce bias include concealment of allocation, blinding of participants, healthcare providers,

investigators, outcome assessors, and statisticians, as well as limiting loss to follow-up.

In critically appraising the ESCAPE trial, you note participants were randomized to their intervention group using a computer-generated variable block randomization scheme, stratified by age and hospital [1]. The use of a computer-generated randomization scheme ensures that allocation is concealed as participants are automatically allocated to an intervention without the influence of researchers or participants. Looking at Table 13.1, you notice that participants allocated to PT were more obese, had more mechanical complaints, more complex degenerative disease, and fewer horizontal tears. This may place the PT group at prognostic disadvantage with regards to functional improvement vs. the surgical group.

With regard to blinding, the participants, physicians, and physical therapists were not blinded due to the nature of the interventions. However, the statistician performing the statistical analysis of the data was blinded [1]. The authors reported 10% loss to follow-up: 18 in the APM arm (11%) and 14 in the PT arm (9%). This modest loss to follow-up provides reassurance that prognostic balance between arms was likely to have been maintained over the course of the trial. Aside from unblinding, you conclude that the ESCAPE study is at low risk of bias.

Did the Investigators Analyze Participants According to the Surgical Treatment they Received, As Well as the Groups to which they Were Assigned?

While randomization aims to balance prognostic factors, participants may not receive their allocated treatment. In the ESCAPE trial, 8 of 159 participants (5%) allocated to APM did not receive surgery. Four of these participants received PT while the other four did not. In the PT group, 47 of 162 participants (29%) received APM, and 17 did not complete all 16 sessions of the PT protocol [1].

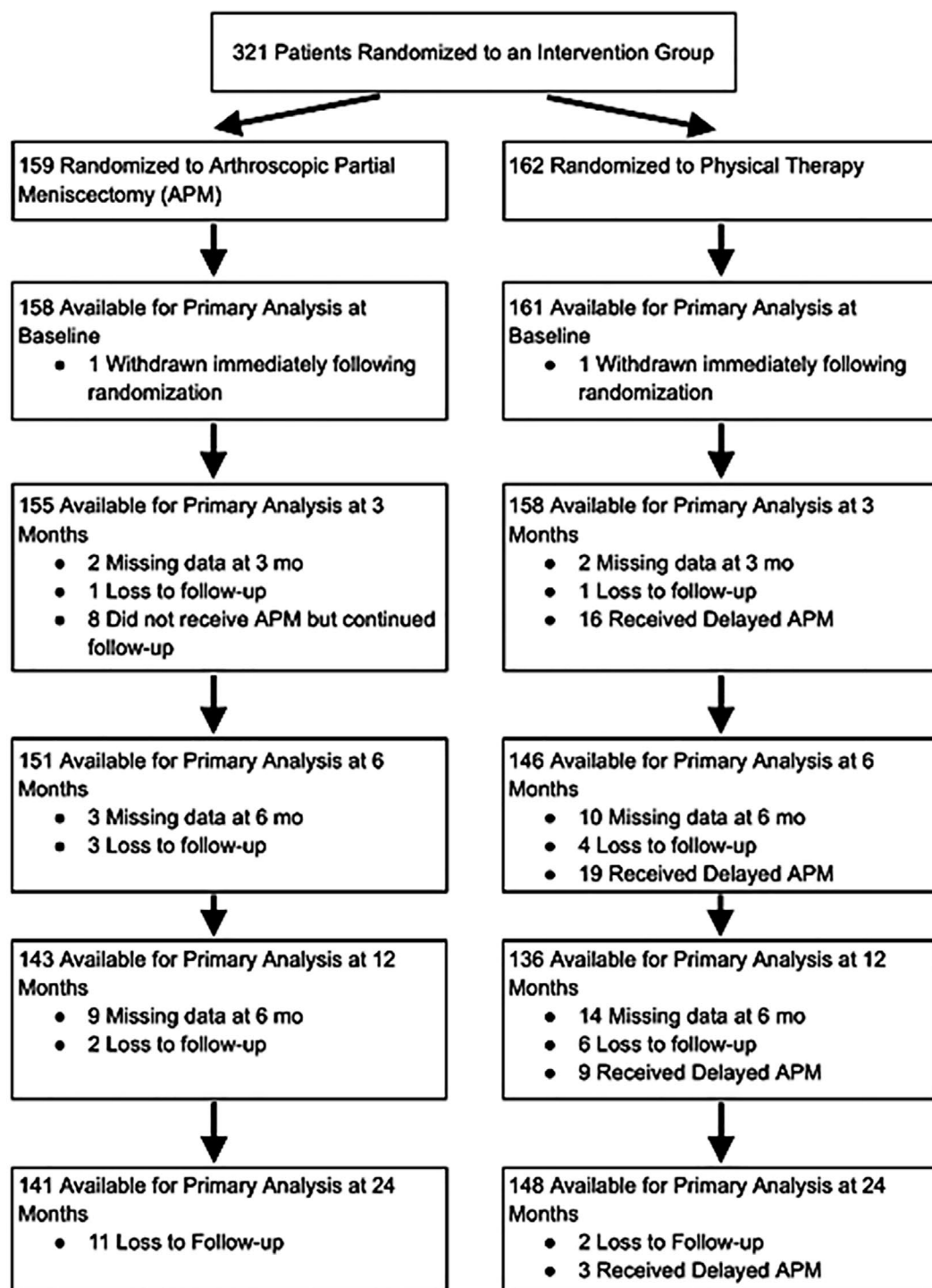


Fig. 13.2 Flow of participants of the ESCAPE trial. (Adapted from van de Graaf et al. [1])

Table 13.1 Baseline characteristics of participants

	Number (%) APM group (<i>n</i> = 158)	PT group (<i>n</i> = 161)
<i>Demographics</i>		
Age, mean (SD)	57.6 (6.5)	57.3 (6.8)
Women	80 (50.6)	81 (50.3)
BMI, mean (SD)	26.7 (3.8)	27.2 (4.0)
18.5 < BMI < 25	56 (35.4)	53 (32.9)
25 ≤ BMI < 30	72 (45.6)	67 (41.6)
30 ≤ BMI < 35	30 (19.0)	41 (25.5)
Mechanical complaints	56 (35.4)	67 (41.6)
<i>Imaging</i>		
Affected meniscus		
Medial	126 (79.7)	136 (84.5)
Lateral	30 (19.0)	25 (15.5)
Both	2 (1.3)	0 (0)
Type of tear on MRI (ISAKOS)	<i>n</i> = 151	<i>n</i> = 152
Longitudinal-vertical	5 (3.3)	5 (3.3)
Horizontal	80 (53.0)	69 (45.4)
Complex degenerative	47 (31.1)	58 (38.1)
Radial	13 (8.6)	10 (6.6)
Vertical flap	2 (1.3)	5 (3.3)
Unclassified	1 (0.7)	5 (3.3)
Horizontal flap	3 (2.0)	0 (0)
Osteoarthritis score on radiographs (KL classification)	<i>n</i> = 150	<i>n</i> = 149
0: None	18 (12.0)	15 (10.1)
1: Doubtful	81 (54.0)	74 (49.7)
2: Minimal	45 (30.0)	55 (36.9)
3: Moderate	6 (4.0)	5 (3.3)
<i>Knee function and pain</i>		
IKDC score (0 [most limitations] to 100 [no limitations]), mean (SD)	44.8 (16.6)	46.5 (14.6)
Knee pain	<i>n</i> = 145	<i>n</i> = 151
VAS during weight-bearing (0 [no pain] to 100 [worst pain]), median (IQR)	61.1 (44.9–83.4)	59.3 (44.9–77.4)

Adapted from van de Graaf et al. [1]

Trialists can address this issue by only analyzing those participants who adhered to the study protocol and received the treatment they were supposed to (a per-protocol analysis); however, this assumes that participants who deviated from protocol were prognostically similar to those who did not—an assumption that is often not met, as nonadherent participants are often sicker and destined to do worse. As such, per-protocol analyses are likely to lead to an overestimation of treatment effects in superiority trials. Using an intention-to-treat (ITT) analysis for superiority trials, where participants are analyzed according to the group to which they were randomized, pro-

vides a more conservative estimate of treatment effects [11].

In noninferiority trials, however, ITT analysis may lead to a misleading conclusion of non-inferiority. Consider a noninferiority trial in which the novel treatment is inferior, and participants are less adherent to the standard treatment. An ITT analysis that included all nonadherent participants would then reduce the apparent effectiveness of the standard therapy and facilitate the misleading conclusion of non-inferiority in comparison with the new approach. As such, a per-protocol analysis is the more conservative approach for noninferiority trials. Investigators

that present both ITT and pre-protocol analyses provide further confidence in a conclusion of noninferiority when the results between both approaches are concordant. The ESCAPE trial reported both an ITT and per-protocol analysis and found similar results.

Did the Investigators Report a Predefined Noninferiority Margin and Guard Against the Unwarranted Conclusion of Noninferiority?

Noninferiority trials are undertaken on the basis that a new approach is either safer, less costly, or more readily available, and the issue is not whether the new approach is more effective than standard care, but only that it is similarly effective (not much worse). Selecting the noninferiority margin (the value that will be acceptable to define “not much worse”) is therefore critical. As noted above, the fixed-margin approach uses 50% of the lower boundary of the 95% CI associated with the treatment effect of the comparator treatment; however, the authors of the ESCAPE trial used a different approach.

To justify their noninferiority margin, the ESCAPE trial wanted to ensure that the comparative effect of PT to APM did not include the possibility that surgery could result in improvements to knee function that patients would consider important. They acknowledged that a minimally important difference (MID) had not been established for their primary outcome—the 100-point International Knee Documentation Committee (IKDC) questionnaire (higher scores are better), among patients with meniscal tears. As such, they used a minimum detectable change of 8.8 points which they (conservatively) rounded down to 8 points [12]. Their noninferiority margin was predefined in their clinical trial registration. Although the minimal detectable change was taken from a study of patients that were different than those enrolled in the ESCAPE trial (older patients, multiple surgical procedures for meniscal injuries, and lack of confirmation of osteoarthritis), studies published after the ESCAPE trial have reported the MID for the IKDC among other

knee surgery populations to be between 9 and 10 points [13, 14].

The selection of a noninferiority threshold based on patient-importance has intuitive appeal compared to approaches based on statistical precision, such as the fixed-margin approach. A recent study found that 72% of cardiovascular noninferiority trials used a noninferiority margin that exceeded a threshold that most patients would consider important [15]. This suggests that clinicians should focus on study results and pay little attention to noninferiority margins.

While the ESCAPE trial’s selection of their noninferiority margin appropriately focusses on patient-important benefit, it implicitly assumes that the average treatment effect on knee function will be experienced by all participants. In reality, there will be a distribution of responses around the mean, with some patients doing better than average and some doing worse. We have evidence from other chronic pain populations that when the average effect is $\frac{1}{2}$ the MID, approximately 10% of patients will still achieve important benefits [16, 17]. As such, the ESCAPE trial could have selected a lower noninferiority margin to increase confidence about noninferiority.

Did the Investigators Power the Study for a Test of Noninferiority?

Prior to conducting a trial, the author should calculate a sample size based on a predefined desired power and significance level, the anticipated effect size, and consideration of practical constraints such as loss to follow-up. The ESCAPE trial used a two-sided alpha value of 0.05, a power of 90%, a standard deviation of 18 points on the IKDC, and a noninferiority margin of 8 points [1]. While considering a loss to follow-up of 20%, their calculations found that 160 participants were required in each intervention group. Noninferiority was demonstrated when the 97.5% CI did not exceed the noninferiority margin.

There were 158 patients enrolled into the APM group and 161 into the PT arm. Loss to follow-up overall was only 10%; however, in the

as-treated analysis, participants were analyzed in three groups: (1) participants allocated to the APM group who received APM, (2) participants allocated to PT who completed 16 sessions without having APM during the follow-up period (<16 sessions was considered a protocol violation), and (3) participants randomized to PT who had APM during follow-up (delayed APM group). Patients who were randomized to APM but did not have surgery, and patients who were randomized to PT but did not complete 16 sessions and did not have delayed APM, were removed from the as-treated analysis. This meant that the PT group was smaller than the sample size calculation requirements and was at risk of being underpowered for a test of noninferiority.

Step 2: What Are the Results?

In the ESCAPE trial, the primary outcome was the change in patient-reported knee function on the Subjective Knee Form of the IKDC, from baseline over 24 months. The IKDC assesses knee function, symptoms, and ability to engage in sports activities, and has been shown reliable and valid for patients with a variety of knee disorders [1, 18]. Secondary outcomes included knee pain (100 mm visual analog scale), general health

(RAND-36 Physical Component Score), osteoarthritis progression (Kellgren-Lawrence classification), and activity level (Tegner Activity Scale).

The ESCAPE trial randomized 321 patients (mean age, 58 [SD 6.6] years; 161 women [50%]), 90% completed the trial, with 29% of the PT group undergoing APM during the 24-month follow-up (Table 13.1). Knee function improved in both groups at all time points, with a between-group difference for the as-treated analysis of 1.7 points on the 100-point IKDC favoring APM (97.5% CI, $-\infty$ to 5.1; *p* value for noninferiority <0.001). These results demonstrate noninferiority as the upper bound of the 97.5% CI falls below the noninferiority margin of 8 points, allowing for rejection of the null hypothesis (Table 13.2).

Among the secondary endpoints, the intention-to-treat analysis found a small overall between-group difference that favored PT over APM for knee pain during weight-bearing (−5.9 mm on the 100 mm VAS, 95% CI −10.3 to −1.4). Intention-to-treat analyses found no significant differences between APM and PT for general health (*p* = 0.08), osteoarthritis progress (*p* = 0.18), or activity level (*p* = 0.73). There were 9 serious adverse events in the APM group and 8 in the PT group. Nonserious adverse events (e.g., reactive arthritis, joint pain resulting in extra consultation or surgical site infection)

Table 13.2 Unadjusted intervention effects for knee function, as measured by the IKDC score (as-treated analysis)

Outcome	APM		Delayed APM	PT		Between-group difference, APM vs. PT (97.5% CI)	<i>P</i> value for noninferiority
	IKDC points	No. of patients (%)	IKDC points	No. of patients (%)	IKDC points	No. of patients (%)	
Baseline	43.9	150 (99%)	40.8	47 (100%)	48.6	97 (100%)	
Overall ^a	66.2	561	57.6	163	66.6	372	1.7 ($-\infty$ to 5.1)
3 months	59.3	147 (97%)	49.8	47 (100%)	64.3	158 (100%)	−2.9 ($-\infty$ to 1.3)
6 months	64.5	144 (95%)	55.7	40 (85%)	65.3	146 (85%)	1.6 ($-\infty$ to 5.7)
12 months	70.5	137 (91%)	63.8	35 (74%)	67.5	136 (74%)	4.9 ($-\infty$ to 9.1)
24 months	71.1	133 (88%)	63.0	41 (87%)	69.2	148 (87%)	4.1 ($-\infty$ to 8.3)

^aOverall estimate over 24 months refers to the overall IKDC score within each group and between groups including all time points. The overall effect is based on a model including intervention group and baseline IKDC score. The No. represents the number of observations entered into the model. Table adapted from van de Graaf et al. [1]

occurred in 9 participants in the APM group and 4 in the PT group.

Were all Patient-Important Outcomes Considered?

The Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) has established 9 core outcomes that should be reported by interventional trials for chronic pain, including adherence (loss to follow-up) and global rating of change; however, asking patients how much they have improved is highly subject to recall bias. The seven remaining outcomes recommended by IMMPACT are pain, physical function, sleep, emotional function, social function, role function, and adverse events [19, 20]. The ESCAPE trial measured pain (100 mm VAS), knee-specific physical function (IKDC), general physical function (RAND-36 Physical Component Score), role function (Tegner Activity Scale), and adverse events. They did not assess the impact of treatment on sleep, emotional function, or social function. As a result, the findings of this trial do not inform the comparative effectiveness of APM vs PT on all patient-important outcomes.

Were the Results Precise?

For their primary outcome, knee-specific physical function as measured by the 100-point IKDC Score of Knee Function, the overall effect over 24-months follow-up was 1.7 points in favor of APM, with an upper boundary of the associated 97.5% CI of 5.1 points. This is below the noninferiority threshold of 8 points, and to increase confidence that an important difference was not missed, they used a 97.5% CI instead of a conventional 95% CI.

You note that the lower boundary of the confidence interval is “infinite,” which intuitively seems imprecise; however, this is the result of the trial authors’ 1-sided test which is designed to explore the null hypothesis that the primary endpoint for PT is worse than APM by 8 points or more on the IKCD. The alternative hypothesis is that PT is not worse than APM by 8 points or

more on the IKCD. As the 97.5% CI demonstrates that the true mean difference is ≤ 5.1 points, the null hypothesis is rejected. As such, the results for the overall assessment of knee function are sufficiently precise; however, when analyzed at individual time points, the difference in IKCD scores is sufficiently precise to exclude the noninferiority margin at 3 and 6 months, but not at 12 months (97.5% CI $-\infty$ to 9.1) or 24 months (97.5% CI $-\infty$ to 8.3).

Were the Investigators Appropriately Interpreting the Concept of Noninferiority?

With regard to the primary outcome of knee function (including all time points), the one-sided 97.5% CI for the as-treated analysis had an upper boundary of 5.1 points on the 100-point IKDC score [1]. This suggests that PT could be considered at most 5.1 points worse than APM in terms of knee function improvement. Since the upper bound of the CI falls below the predefined noninferiority margin of 8 points, it suggests that PT is noninferior to APM in terms of knee function improvement. Thus, the null hypothesis was rejected in favor of the alternative hypothesis. The authors of the article appropriately interpreted these results: “This multicenter RCT showed that, in patients older than 45 years old with knee pain and nonobstructive meniscal tears, PT was noninferior to APM for knee function over a 24-month follow-up period.” Further, they acknowledged that noninferiority was not confirmed at 12 and 24-months follow-up.

Step 3: How Can I Apply the Results to My Patient or My Clinical Practice?

Were the Participants Similar to My Patient?

Your patient is similar to the baseline demographics provided in Table 1 of the ESCAPE study [1]. If your patient had been approached

by the ESCAPE trial, you are confident that she would have met inclusion criteria. You conclude that your patient is similar to the study participants in the ESCAPE trial and that results would apply to her.

Are the Likely Advantages of PT Worth the Potential Harm and Costs?

The ESCAPE trial found that PT was noninferior to APM for knee function over a 24-month period; however, this was not confirmed at the individual time points of 12- and 24-month follow-up. PT was found to provide a small improvement in pain upon weight-bearing over APM (ITT analysis mean difference -5.9 mm on a 100 mm VAS), and no difference in general health, osteoarthritis progress, or activity level. Adverse events were similar between group.

PT as delivered in the ESCAPE trial requires a commitment by patients to attend biweekly, 30-min sessions for 2 months, whereas surgery requires a single session with a recovery period. The overall cost of surgery is higher than PT; however, who bears the costs will depend on the jurisdiction and possibly on whether patients have health insurance. Further, PT delivered in the ESCAPE trial consisted of cardiovascular, coordination/balance, and closed kinetic chain strength exercises. In fee-for-service healthcare systems, there are greater financial incentives for physiotherapists to provide passive modalities (e.g., ultrasound, low-level laser, transcutaneous electrical nerve stimulation) versus active rehabilitation. Further, in the ESCAPE trial, almost 1 in 3 participants randomized to PT did not achieve satisfactory results and elected to receive surgery.

Given the greater burden of 2 months of PT over APM, possibly greater costs, similar outcomes, and the possibility of receiving surgery after a full course of PT, the decision to pursue one or the other is dependent on patient's values and preferences. One patient who was concerned about rare, but serious harms of surgery, and keen to try active rehabilitation first and then surgery if unsuccessful, may prefer PT. Another who did not have easy access to PT, would have to bear

costs of rehabilitation out-of-pocket, and who was not concerned about possible harms of surgery after being appraised of the evidence, may elect to pursue APM.

Conclusion

Noninferiority trials are an appropriate study design to explore whether a novel approach that has advantages in cost, harms, or availability is similarly effective to the existing standard. Criteria for assessing validity include those used for conventional RCTs, but also include consideration of whether the conventional approach was delivered optimally and if findings are robust to both an ITT and per-protocol analysis. Moreover, clinicians, in consultation with their patients, should determine whether the loss in effectiveness suggested by the upper boundary of the confidence interval associated with the treatment effect of the novel approach is acceptable, regardless of a specific trial's noninferiority threshold.

Appendix A: Literature Search

Database: PubMed (<https://pubmed.ncbi.nlm.nih.gov/>)

Date Searched: April 20, 2024

of results: 4

Search Strategy:

- 1. "nonobstructive meniscal tears"
- 2. "arthroscopic partial meniscectomy"
- 3. "physical therapy"
- 4. 1 AND 2 AND 3

References

1. van de Graaf VA, Noorduyn JCA, Willigenburg NW, Butter IK, de Gast A, Mol BW, Saris DBF, Twisk JWR, Poolman RW. Effect of early surgery vs physical therapy on knee function among patients with nonobstructive meniscal tears:

- the ESCAPE randomized clinical trial. *JAMA*. 2018;320(13):1328–37.
2. Hariton E, Locascio JJ. Randomised controlled trials – the gold standard for effectiveness research: study design: randomised controlled trials. *BJOG*. 2018;125(13):1716.
 3. Phillips MR, Kaiser P, Thabane L, Bhandari M, Chaudhary V, Wykoff CC, Sivaprasad S, Sarraf D, Bakri SJ, Garg SJ, Singh RP, Holz FG, Wong TY. Risk of bias: why measure it, and how? *Eye*. 2022;36(2):346–8.
 4. Christensen E. Methodology of superiority vs. equivalence trials and non-inferiority trials. *J Hepatol*. 2007;46(5):947–54.
 5. Wang B, Wang H, Tu XM, Feng C. Comparisons of superiority, non-inferiority, and equivalence trials. *Shanghai Arch Psychiatry*. 2017;29(6):385–8.
 6. U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Non-inferiority clinical trials to establish effectiveness: guidance for industry. FDA; 2020, April 27. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/non-inferiority-clinical-trials>.
 7. Althunian TA, de Boer A, Groenwold RHH, Klungel OH. Defining the noninferiority margin and analysing noninferiority: an overview. *Br J Clin Pharmacol*. 2017;83(8):1636–42. <https://doi.org/10.1111/bcp.13280>.
 8. Piaggio G, Elbourne DR, Pocock SJ, SJW E, Altman DG, CONSORT Group. Reporting of non-inferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA*. 2012;308(24):2594–604. <https://doi.org/10.1001/jama.2012.87802>.
 9. Aberegg SK, Hersh AM, Samore MH. Empirical consequences of current recommendations for the design and interpretation of noninferiority trials. *J Gen Intern Med*. 2018;33(1):88–96. <https://doi.org/10.1007/s11606-017-4161-4>.
 10. Thoma A, Farrokhhyar F, Waltho D, Braga LH, Sprague S, Goldsmith CH. Users' guide to the surgical literature: how to assess a noninferiority trial. *Can J Surg*. 2017;60(6):426–32. <https://doi.org/10.1503/cjs.000317>.
 11. Gupta SK. Intention-to-treat concept: a review. *Perspect Clin Res*. 2011;2(3):109–12. <https://doi.org/10.4103/2229-3485.83221>.
 12. Crawford K, Briggs KK, Rodkey WG, Steadman JR. Reliability, validity, and responsiveness of the IKDC score for meniscus injuries of the knee. *Arthroscopy*. 2007;23(8):839–44. <https://doi.org/10.1016/j.arthro.2007.02.005>.
 13. Ogura T, Ackermann J, Mestriner AB, Merkely G, Gomoll AH. The minimal clinically important difference and substantial clinical benefit in the patient-reported outcome measures of patients undergoing Osteochondral allograft transplantation in the knee. *Cartilage*. 2021;12(1):42–50. <https://doi.org/10.1177/1947603518812552>.
 14. Agarwalla A, Puzziello RN, Liu JN, Cvetanovich GL, Gowd AK, Verma NN, Cole BJ, Forsythe B. Timeline for maximal subjective outcome improvement after anterior cruciate ligament reconstruction. *Am J Sports Med*. 2019;47(10):2501–9. <https://doi.org/10.1177/0363546518803365>.
 15. Ofori S, Cafaro T, Devereaux PJ, Marcucci M, Mbuagbaw L, Thabane L, Guyatt G. Noninferiority margins exceed superiority effect estimates for mortality in cardiovascular trials in high-impact journals. *J Clin Epidemiol*. 2023;161:20–7. <https://doi.org/10.1016/j.jclinepi.2023.06.022>.
 16. Busse JW, Wang L, Kamaleldin M, Craigie S, Riva JJ, Montoya T, Mulla SM, Lopes LC, Vogel N, Chen E, Kirmayr K, De Oliveira K, Olivieri L, Kaushal A, Chaparro LE, Oyberman I, Agarwal A, Couban R, Tsoi L, Lam T, Vandvik PO, Hsu S, Bala MM, Schandelmaier S, Scheidecker A, Ebrahim S, Ashoorion V, Rehman Y, Hong PJ, Ross S, Johnston BC, Kunz R, Sun X, Buckley N, Sessler DI, Guyatt GH. Opioids for chronic noncancer pain: a systematic review and meta-analysis. *JAMA*. 2018;320(23):2448–60. <https://doi.org/10.1001/jama.2018.18472>.
 17. Wang L, Hong PJ, May C, Rehman Y, Oparin Y, Hong CJ, Hong BY, AminiLari M, Gallo L, Kaushal A, Craigie S, Couban RJ, Kum E, Shanthanna H, Price I, Upadhye S, Ware MA, Campbell F, Buchbinder R, Agoritsas T, Busse JW. Medical cannabis or cannabinoids for chronic non-cancer and cancer related pain: a systematic review and meta-analysis of randomised clinical trials. *BMJ (Clin Res Ed)*. 2021;374:n1034. <https://doi.org/10.1136/bmj.n1034>.
 18. Higgins LD, Taylor MK, Park D, Ghodadra N, Marchant M, Pietrobon R, Cook C, International Knee Documentation Committee. Reliability and validity of the international knee documentation committee (IKDC) subjective knee form. *Joint Bone Spine*. 2007;74(6):594–9. <https://doi.org/10.1016/j.jbspin.2007.01.036>.
 19. Turk DC, Dworkin RH, Allen RR, Bellamy N, Brandenburg N, Carr DB, Cleeland C, Dionne R, Farrar JT, Galer BS, Hewitt DJ, Jadad AR, Katz NP, Kramer LD, Manning DC, McCormick CG, McDermott MP, McGrath P, Quessy S, Rappaport BA, Robinson JP, Royal MA, Simon L, Stauffer JW, Stein W, Tollett J, Witter J. Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. *Pain*. 2003;106(3):337–45.
 20. Turk DC, Dworkin RH, Revicki D, Harding G, Burke LB, Cella D, Cleeland CS, Cowan P, Farrar JT, Hertz S, Max MB, Rappaport BA. Identifying important outcome domains for chronic pain clinical trials. *Pain*. 2008;137(2):276–85.

Expertise-Based Randomized Controlled Trials

14

Tyler McKechnie, Daniel Waltho, Kristen Davidge,
and Cagla Eskicioglu

Abbreviations

ANOVA	Analysis of Variance
ASES	American Shoulder and Elbow Surgeons
CONSORT	Consolidate Standards of Reporting Trials
EBRCT	Expertise-Based Randomized Controlled Trial
ITT	Intention-to-Treat
PICOT	Population, Intervention, Comparator, Outcome(s), Time
PRO	Patient Reported Outcome
RCT	Randomized Controlled Trial
USD	United States Dollars
WOSI	Western Ontario Shoulder Instability Index

Clinical Scenario

You are an orthopedic surgeon in the community seeing a 25-year-old, male, right-hand dominant, hockey player in your clinic. He has been diagnosed with recurrent, traumatic, anterior glenohumeral instability of his right shoulder, which followed a hockey-related injury 5 years ago. The patient reports multiple episodes of his right shoulder “popping out” since the injury (approximately two episodes per year). Being quite proficient with the open repair technique for shoulder stabilization, you offer this procedure to the patient. However, the patient remarks that a friend of his with the same condition was treated at a tertiary care hospital with an arthroscopic repair of his shoulder. Your patient asks you which of the two procedures will result in a better overall result in terms of his quality of life. Having only performed the open repair in your practice, you are unsure of the answer and so you consult the literature.

T. McKechnie · C. Eskicioglu (✉)
Division of General Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: tyler.mckechnie@medportal.ca;
eskicio@mcmaster.ca

D. Waltho
Department of Surgery, Division of Plastic Surgery,
Oakville-Trafalgar Memorial Hospital,
Oakville, ON, Canada
e-mail: daniel.waltho@medportal.ca

K. Davidge
Division of Plastic Surgery, Department of Surgery,
University of Toronto, Toronto, ON, Canada
e-mail: kristen.davidge@sickkids.ca

Introduction

The randomized controlled trial (RCT) is an experimental study design, which randomly assigns participants typically into two arms: experimental group (typically a novel intervention) and a control group (typically the current standard intervention). Ideally, the intervention(s) and comparator(s) are carried out by the same party to ensure consistency and standardization between

groups (see Chap. 11). However, in certain cases, these treatments differ significantly in terms of the skills required to implement them. This issue represents a large factor dissuading nonpharmacological RCTs. RCTs involving surgical intervention(s) are one such example, wherein surgeons with a specific expertise and skill set may need to be selectively involved in the treatment of one group. These circumstances represent a challenge when conducting an RCT of this nature.

Novel surgical interventions can have a learning curve that may be difficult for all participating surgeons to overcome. To address this dilemma, an expertise-based RCT (EBRCT) design has been proposed, wherein two unique sets of health professionals possessing a relatively similar expertise with one of the study treatments are assigned to that respective intervention group. This method was first proposed in 1980 by van der Linden and has been applied sparingly since then [1].

This design allows for comparable skill in executing the therapies in question and may also lend itself to improved participation and compliance amongst surgeons. As such, EBRCTs have emerged as an effective study design in the surgical literature and remain the standard for RCTs comparing surgery versus medical management or comparing two or more distinct surgical techniques [2]. Indeed, their use in surgical settings has been increasing over the past decade [2]. It is prudent for surgeons to be familiar with this design when attempting to answer their clinical or research questions. The surgeon, however, must exercise caution when interpreting the results of an EBRCT. Careful consideration of the methodology used in these RCTs is critical to making appropriate clinical decisions.

This chapter provides the readers with the tools to appraise EBRCTs in surgery. The framework for this appraisal is found in Box 14.1. Because the EBRCT is a modified RCT, some of the appraisal items in Box 14.1 are similar to those found in Chap. 11. More attention, however, will be spent on the differences between the two study designs.

Box 14.1 Guidelines of How to Assess an Expertise-Based RCT

A. *Are the results valid?*

- (i) Was the learning curve taken into consideration?
 - (a) Sufficient level of expertise in each treatment group objectively determined and clearly stated
 - (b) Matching of equally skilled surgeons between each treatment group
 - (c) Sufficient number of surgeons in each treatment group
 - (d) Sufficient level of expertise in each treatment group objectively determined and clearly stated
- (ii) Were the participants randomized?
- (iii) Was the randomization concealed?
- (iv) Were the experimental and control group similar in terms of prognostic factors?
- (v) Was the randomization stratified?
- (vi) Were participants analysed in the group they were initially randomized into at enrolment?
- (vii) Was follow-up complete?

B. *What are the results?*

- (i) How were the results of the trial being measured?
- (ii) How large is the treatment effect?

C. *Are the results applicable to my patients?*

- (i) Does the study represent my patient population?
- (ii) Does the trial consider all relevant patient-important outcomes?
- (iii) Do the benefits of the procedure outweigh any potential risks and costs?

Literature Search

To identify the best evidence and inform our patient a literature search was performed according to the Users’ Guide for Surgical Literature: How to perform a high-quality literature search [3]. In this case, the best level of evidence would be a high-quality RCT or preferably a meta-analysis thereof. Designing our research question, using the PICOT format (Box 14.2), allows us to choose important key words for our search. Your clinical question relevant to this scenario would be as follows.

Box 14.2 Stating the Research Question Using the PICOT Format

Population: males with recurrent, traumatic, anterior glenohumeral instability

Intervention: open shoulder repair

Comparison: arthroscopic shoulder repair

Outcome: disease or condition-specific quality of life

Time Horizon: long-term postoperative period (up to 2 years)

Using the PICOT format above, the research question would look like “What is the quality of life of males with recurrent, traumatic, anterior glenohumeral instability after an open shoulder repair, as compared to those who receive an arthroscopic shoulder repair?” To answer your research question, you search PubMed Clinical Queries with the search terms “glenohumeral instability” AND “open” AND “arthroscopic” AND “quality of life” (Appendix 1). The search aimed to include all studies from database inception up until March 1, 2024. You identify an RCT, entitled “A Randomized Clinical Trial Comparing Open and Arthroscopic Stabilization for Recurrent Traumatic Anterior Shoulder Instability: 2-year Follow-Up with Disease-Specific Quality-of-Life Outcomes,” by Mohtadi et al. [4]. In this RCT, an expertise-based methodology was used. Patients were randomized to either open or arthroscopic shoulder repairs, and further randomized to a surgeon with sufficient experience in that chosen technique. Important to this design, surgeons in the trial were assigned to their preferred technique a priori. Key methodological features of this study are outlined in Table 14.1.

Table 14.1 Key methodological features of the Mohtadi et al. [4] article

Study period	2001–2008
Source of sample	University of Calgary Sports Medicine Centre
Source of funding	Calgary Orthopaedic Research and Education Fund Calgary Regional Health Authority Research and Development Fund Hip Hip Hooray! (Canadian Orthopaedic Foundation)
Sample size + Results	Open repair: $n=98$ Arthroscopic repair: $n=98$ Statistically significant increase in % of cases with dominant shoulder involvement Statistically significant increase in mean time from injury to repair Age, sex, contact sport involvement, dislocations and # anchors used statistically not significant
Analysis	Independent sample t tests were used to compare mean quality of life scores groups at 2 years post-operatively Adjusted Bonferroni comparisons and repeated-measures analyses of mean quality of life scores were conducted with use of a mixed-model analysis of variance (ANOVA) for treatment group and time of assessment Intention to treat was utilized for all patients Chi-square analysis was used to compare recurrence rates between groups Logistic regression analysis was performed to assess the independent predictors of recurrence Sensitivity analysis was performed on loss to follow-up/withdrawal

Are the Results Valid?

In this section, we will determine whether the methodology used for this EBRCT was robust, such that we can be confident that any differences, or lack thereof, between treatment groups are more likely to be true. Thus, we will be asking the following questions:

- (i) Was the learning curve taken into consideration?
- (ii) Were the participants randomized?
- (iii) Was the randomization concealed?
- (iv) Were the experimental and control group similar in terms of prognostic factors?
- (v) Was the randomization stratified?
- (vi) Were participants analyzed in the group they were initially randomized into at enrolment?
- (vii) Was follow-up complete?

Was the Learning Curve Taken into Consideration?

This aspect of validity is a common pitfall in surgical RCTs and is a primary focus of optimization in the EBRCT. When an RCT comparing two surgical interventions is performed without an expertise-based design, the experimental group is often a relatively novel procedure, with which many surgeons may not have sufficient experience. The presence of a learning curve will likely create a discrepancy between the surgeon's proficiency and their ultimate degree of success with the newer surgical treatment. This discrepancy may translate into a systematic difference in outcomes between groups favouring the older procedure. The EBRCT design assigns patients to an intervention group where only surgeons who are well experienced in that specific surgical procedure perform the intervention. Further this design can match patients based on surgeon's level of experience. In other words, if one patient randomized to the experimental treatment arm receives their respective procedure by a surgeon with a certain degree of experience in that technique, he/she will be matched to a patient randomized to the com-

parison arm who receives their respective procedure from another surgeon with a similar degree of experience in the comparison technique. Therefore, this design allows the surgeons to perform their preferred surgical technique, which in turn increases validity and feasibility of the trial.

To ensure validity in this category, one must ensure that the expertise-based protocol was carried out appropriately. Therefore, this chapter proposes the following sub-criteria that must be fulfilled to determine validity:

- (a) Sufficient level of expertise in each treatment group objectively determined and clearly stated.
- (b) Matching of equally skilled surgeons between each treatment group.
- (c) Sufficient number of surgeons in each treatment group.
- (a) *Sufficient level of expertise in each treatment group objectively determined and clearly stated*

Level of surgeon expertise is at the crux of this RCT design. To conclude that the trial results are valid, it must first be determined that the surgeons performing the treatment(s) are adequately trained, such that resultant outcomes (i.e. Success, complications, etc.) are not attributable to poor surgical technique. Therefore, the trial should provide an objective metric for each surgeon based on expertise in their assigned operative treatment. There is evidence to suggest that for many procedures, at least 5 years of experience is required to reach the plateau of the learning curve [2]. In the Mohtadi et al. [4] trial, one surgeon in each group had 10 years of experience, while the remaining surgeons in each group had 2–5 years of experience. No sensitivity analysis was done on the study outcomes based on expertise level. Moreover, the authors are assuming that each surgeon has performed a similar number of cases in their respective technique. Whereas a surgeon could, in theory, have performed only one open procedure a year for the past 10 years and another surgeon could have performed 100 arthroscopic procedures a year for 10 years. A more appropriate metric for measur-

ing expertise may be the total number of a particular procedure performed.

(b) *Matching of equally skilled surgeons between each treatment group*

Another critical component of the EBRCT protocol is matching of surgeons based upon their level of expertise in their respective treatment group. In the above sub-criterion, we identify a diversity of skill level with the surgeons involved in the trial. While it is important to enlist surgeons in all treatment groups who have sufficient training to be considered “experts” in a given surgical procedure, these surgeons still exist on a spectrum in terms of how seasoned they are in that technique. Therefore, further matching of surgeons based on their level of expertise is indicated to control for any variation in outcomes that may be attributed to how many of these procedures a surgeon has performed. In the Mohtadi et al. [4] trial, surgeons were matched by experience based on the approximate number of years they were performing their preferred technique. This ensures that for every patient randomized to a surgeon who has performed the arthroscopic technique for a certain number of years, a patient would also be randomized to a surgeon who has performed the open technique for the same number of years. However, as mentioned above, the number of years may not necessarily correspond to the number of procedures performed and matching would be more suitably performed based on the latter metric.

(c) *Sufficient number of surgeons in each treatment group*

To further optimize the execution of the EBRCT design, the trial must ensure an adequate number of surgeons in each treatment group, thereby limiting the number of incidences where patients may cross over to the other group. The number of surgeons required is contingent upon both the volume and priority of cases. There is likely to be increased difficulty with respect to the availability of any surgeon in the desired group when the cases are frequent and need to be

done on an emergent basis. In the case of the Mohtadi et al. [4] trial, cases were done on an elective basis and a total of 226 eligible patients were included in expertise-based randomization between 2001 and 2008. Throughout the trial period, two surgeons performed open repairs and three surgeons performed arthroscopic repairs. One patient, who was randomized to open repair, crossed over. Therefore, we can conclude that there were an adequate number of surgeons in both groups to support this trial.

Based on these criteria, we can conclude that a fairly robust expertise-based design was carried out throughout the trial; however, caution must be taken when interpreting the results as true expertise of the surgeons involved is not clearly reported.

Were the Participants Randomized?

In this trial, the allocation was determined using computer-generated, variable-block-size randomization. Block randomization is simply randomization within blocks proportional to the allocation ratio used to distribute patients to each treatment arm. Variable-block-sizes decrease predictability in the allocation process by varying the number of patients assigned to a given block throughout the allocation process. In the Mohtadi et al. [4] trial, appropriate randomization is utilized. Because this trial is an expertise-based design, allocation determined both the technique and surgeon as discussed above. Patients would then meet with their assigned surgeon, who discussed the technique, confirmed eligibility, and addressed any concerns. Randomization was, therefore, sufficiently executed in this trial.

Was the Randomization Concealed?

Above all, concealment of allocation should apply to those enrolling a patient into the trial, as it ensures that systematic selection bias associated with decisions to enrol certain patients into each treatment arm is limited. In the Mohtadi et al. [4] RCT, the allocation was concealed with

consecutively numbered opaque envelopes. Assuming no tampering occurred, we can be satisfied that allocation concealment was sufficient for those enrolling patients initially.

Were the Treatment Arms Blinded?

This trial is based upon two significantly different surgical approaches, open versus arthroscopic, which incorporate different incisional patterns, therefore, blinding of patients is difficult to achieve. Mohtadi et al. [4] were unable to sufficiently blind the patients to the treatment selection. Furthermore, the authors state that the research assistant performing the clinical examinations for evaluation of treatments was aware of the treating surgeon. This is due to both the difference in incisional patterns indicated above, as well as the expertise-based nature of surgeon allocation. Finally, while blinding is impossible to achieve for the surgeons providing treatment, it is unlikely to introduce bias in this case. In traditional surgical RCTs, where the same surgeon performs both interventions, there is a concern that if a surgeon strongly believes in the superiority of one of the two interventions, then either consciously, or subconsciously, patients may be treated differently (e.g., obtaining more meticulous hemostasis) or introducing co-interventions. However, the use of the EBRCT study design avoids these concerns especially if this is a pragmatic EBRCT.

Were the Experimental and Control Groups Similar in Terms of Prognostic Factors?

To ensure validity, we need to determine if differences in outcomes between treatment groups may be attributable to differences between patient populations in each group. Randomization attempts to minimize this by limiting systematic biases that may cause certain patients to fall into one of the arms and not the other. In other words, randomization should lead to balanced prognosis between treatment groups. Care should be taken to observe the initial demographic data of each group. In the

Mohtadi et al. [4] trial, inferential statistics were used to compare baseline demographics between treatment arms. However, according to the 2010 Consolidate Standards of Reporting Trials (CONSORT) statement for parallel group randomised trials, significance testing of baseline differences in RCTs should not be performed. Baseline significance testing is not wrong, however it may mislead readers because as long as randomization and concealment are performed appropriately, any difference in baseline variables between treatment arms are simply due to chance or small sample size (which is not the case with Mohtadi et al. [4]), as opposed to bias. Rather, to assess for bias, readers must scrutinize the similarity of the absolute value of the baseline variables and determine whether any observed differences may account for imbalanced prognosis. The authors for the Mohtadi et al. [4] trial reported significant baseline differences in the proportion of dominant shoulders being treated and the length of time from initial injury, and no differences between the groups in other important baseline variables such as age, gender, collision/contact sport involvement, number of previous dislocations, or the number of suture anchors used intraoperatively. Assessing the absolute values of the baseline data and interpreting them in terms of clinical relevance, however, is likely more informative for readers. While a larger proportion of open repairs were done on the patient's dominant shoulder, the clinical importance of this variable is not entirely clear. Perhaps a dominant shoulder repair group may lead those patients to be more motivated or capable of postoperative therapy and recovery or may also cause postoperative injury due to overexertion. Moreover, because this primary outcome was based on a quality of life measurement, perceived improvement in disease-specific quality of life may be greater if surgery is performed on a dominant shoulder injury. Whereas the longer period from injury occurred with the open group may be a clinically relevant baseline discrepancy that has a more direct impact on internal validity. A relative increase in the time from injury with the open repairs could bias the results in favour of the arthroscopic procedure, as a greater time until repair typically yields poorer results. Overall, in

the Mohtadi et al. [4] trial, it is not clear whether this baseline characteristic discrepancy affects validity; however, readers should note that the imbalance of shoulder dominance may bias the results toward the open treatment arm, while the imbalance in terms of time from index injury may bias the results toward the arthroscopic treatment arm. Adequate balance between the remaining baseline variables suggest that internal validity is still fairly well maintained for this trial despite these differences.

Was the Randomization Stratified?

The further subdivision of participants beyond the treatment administered is called stratification. Stratification attempts to ensure equal allocation of subgroups of participants based on a desire for uniform patients between groups as described in the previous section. This exercise, therefore, limits systematic differences that may skew results in favor of one treatment. Stratification can be based on gender, age, or other demographic factors. In keeping with an expertise-based design, in the Mohtadi et al. [4] trial, stratification occurred at the level of the surgeon, wherein surgeons of a similar level of expertise in their respective technique were matched. The authors comment on the randomization of participants being stratified accordingly; however, they do not specify the patient characteristics for which stratification was based upon. Based on the demographic information discussed in the previous section, we can be satisfied with homogeneous strata between both treatment arms. However, these baseline data were not shown by stratum in the paper or this chapter.

Were Participants Analyzed in the Group They Were Initially Randomized into at Enrolment?

Intention-To-Treat (ITT) analysis is a method wherein all patients enrolled and randomly allocated to a treatment are analyzed in that group to which they were randomized, regardless of devi-

ations that occur after randomization (e.g., protocol violations, loss to follow-up/withdrawal, patient noncompliance). This practice optimizes validity by ensuring that the study patients remain similarly distributed between study groups, thus avoiding crossover of participants to a group, to which they were not randomized. In the Mohtadi et al. [4] RCT, all patients were analyzed on an ITT basis so we can conclude that validity has been optimized in this regard. If RCTs are not analyzed according to ITT (i.e., per protocol analysis), missing data and patient crossover risks disrupting prognostic balance and may introduce bias that can threaten internal validity. In the trial by Mohtadi et al. in particular, analyzing the data according to a per-protocol analysis could have significantly biased the findings as 15 patients (13.3%) that were randomized to each arm did not undergo surgery and one patient crossed over from the open repair group to the arthroscopic repair group (1.0%) [4].

Was Follow-Up Complete?

To capture treatment effects, patients must be followed up both frequently and for a sufficient duration. Ideally, multiple patient follow-ups for a patient's lifetime would provide the most accurate data; however, feasibility must be considered. Therefore, the surgeon should identify the duration and frequency of follow-up that would be practical based on the disease and treatment being studied. These follow-up periods usually correspond to the surgeon's typical follow-up routine in their clinical practice. In addition to a sufficient follow-up period, a sufficient proportion of patients in the study should complete follow-up to ensure validity. In the Mohtadi et al. [4] RCT, follow-up occurred at 3 months, 6 months, 1 year, and 2 years postoperatively. The authors report that 19 patients in the open repair group and 14 patients in the arthroscopic repair group were lost to follow-up at 2 years. To optimize validity, a study should aim to have less than 20% of participants lost to follow-up [5]. The number of participants lost to follow-up must also be interpreted in the context of the type of outcome (i.e., continu-

ous, dichotomous, etc.) and the risk of the outcome to correctly interpret the effect of missing data on internal validity. In this case, only 71% of patients initially randomized completed 2-year follow-up (i.e., 29% of participants lost to follow-up). The authors reported that participants missing at 2-year follow-up did not differ substantially at baseline between groups. Moreover, they conducted a sensitivity analysis comparing per protocol analysis to ITT analysis and found no difference in effect estimate or statistical significance with regard to the primary outcome. Therefore, while the proportion of missing outcome data does raise concern in this RCT, the sensitivity analysis gives some confidence that the internal validity is not seriously compromised.

What Are the Results of the Trial?

Once the validity of the trial has been established, we must determine the magnitude of the treatment effect. In this section, we will further understand the results of the trial by answering the following questions:

- (i) How were the results of the trial being measured?
- (ii) How large is the treatment effect?

How Were the Results of the Trial Being Measured?

Study outcomes should be measured using an objective tool that ideally has been appropriately validated and shown to be reliable. The primary outcome was disease-specific quality of life. The authors utilize a patient-reported outcome (PRO) measure to address this outcome. The Western Ontario Shoulder Instability Index (WOSI) was chosen. It uses a 100-mm visual analog scale response format. Patients completed the WOSI at baseline, at 3 and 6 months postoperatively, and at 1 and 2 years postoperatively. Evaluating the original psychometric study on this PRO, we determine that it is both validated and reliable in this category [6].

The secondary outcome was shoulder function, including both patient-reported and clinician-reported measures. The American Shoulder and Elbow Surgeons (ASES) scale, a shoulder-specific functional assessment tool, is a PRO assessing shoulder function. The score was determined based on patient evaluation of pain, instability, and activities of daily living and utilized a 0–100 scale. Once again, exploring the psychometric literature, we are assured that this outcome is both valid and reliable in the study population [7]. Furthermore, clinical assessment of shoulder range of motion was performed by trained, independent research assistants using a standardized tool (i.e., goniometer). Clinical assessment was presumably also used to measure adverse events/complications and recurrence rates. Based on this article, we can conclude that this method of clinical assessment is both valid and reliable. Recurrent instability was another secondary outcome; however, details surrounding how this was assessed were not disclosed by the authors.

How Large Is the Treatment Effect ?

In our clinical scenario, we ask which of the two treatments is superior. To answer our initial clinical question, we must first determine whether there are statistically significant differences between the two techniques. Assessing the Mohtadi et al. [4] article with respect to the primary outcome, no statistically significant difference was found between the two treatments in terms of disease-specific quality of life (mean WOSI scores). Further, no statistically significant difference was found between the two treatments in terms of shoulder function (ASES scores and active range of motion). However, recurrence rates at 2 years were significantly different, with 11% in the open group and 23% in the arthroscopic group ($p = 0.05$). Adverse events were described descriptively and there were no clinically important differences in adverse events. Based on these data, we can only conclude that a difference between the two procedures exists based on the recurrence of instability.

Are the Results Applicable to Clinical Practice?

According to Guyatt et al., in order to accept that the evidence in a particular study is applicable to your patient(s), three critical questions must be addressed [8]:

- (i) Does the study represent my patient population?
- (ii) Does the trial consider all relevant patient-important outcomes?
- (iii) Do the benefits of the procedure outweigh any potential risks and costs?

Does the Study Represent My Patient Population?

Prior to applying the results of a study to your patient(s), you should review the study population to ensure that the patient demographics are in line with that of your patient(s). The authors presumably included all-comers with recurrent traumatic anterior shoulder instability. However, those patients with glenoid fracture or bone loss seen on preoperative X-ray were excluded. Assessing the demographic breakdown of patients, mean age was 27.8 and 27.2 between the open and arthroscopic group, respectively, with an 82% male predominance in both groups. Based on these fundamental demographics, we can conclude that the study population is representative of our patient from the clinical scenario. Furthermore, approximately half of these patients injured their shoulder in a contact sport, while a third to a half of the shoulder instability in the study involved the dominant shoulder. Finally, shoulder injuries were around the 6-year mark at the time of both treatments. Therefore, we are confident that our patient's clinical scenario is relatively well reflected in the study.

Does the Trial Consider All Relevant Patient-Important Outcomes?

Looking at our clinical scenario, the main outcome in question was quality of life. Indeed, this is likely to be the most important outcome for the

patient, as it provides a practical culmination of symptoms and functional aspects germane to the condition and procedure. The Mohtadi et al. [4] RCT addresses disease-specific quality of life as a primary outcome. In addition, the authors include shoulder function, including recurrent instability, and safety outcomes secondarily. Shoulder function provides appropriate data to confirm the mechanism of any quality of life improvements that may have been found in the study. Safety outcomes, including complications and adverse outcomes, are critical in deciding whether to adopt or endorse a particular surgical intervention. The authors presumably study all complications and adverse outcomes throughout the follow-up period. Overall, the included outcomes are both clinically important and comprehensive in assessing the two treatment arms.

Do the Benefits of the Procedure Outweigh Any Potential Risks and Costs?

Another practical argument for deciding upon a treatment is the risk and cost to the patient, healthcare system, and society. A significant benefit in a particular treatment decision should outweigh potentially undue risks to the patient or excessive costs that may be associated with a novel surgery. There were no significant complications reported within this study. Both treatment arms had a small number of cases of transient nerve dysfunction which resolved spontaneously. The authors did not include a cost analysis within their study. Costs can be dichotomized into direct and indirect and can be further divided into costs to the healthcare system, the patient, and to the society. In this case, if we review the literature, we come across a recent cost-effectiveness study by Min et al. published in 2018 [9] According to this study, the average direct cost of surgery for the arthroscopic shoulder repair was US\$20,385, and the average cost of surgery of the open repair was US\$21,389. The included studies reporting these data were published between 2003 and 2017. The authors conclude that both the arthroscopic and the open procedures are highly cost-effective, but the arthroscopic procedure is

more cost-effective due to a lower health utility state after a failed open procedure [9]. However, this study does not incorporate indirect costs to patient or society and further costs are based on a single institution. Further research would be indicated to determine true risk and benefit ratio; however, there appears to be similarity between the groups so as not to dissuade the surgeon or patient from either procedure.

Conclusion

The EBRCT is a promising answer to the challenges that are faced when applying the RCT design to surgical interventions. By obviating the issue of learning curve, the surgeon can be more confident that differences between treatment groups are not attributable to surgeon inexperience. Furthermore, concerns of surgeons personally believing in the superiority of a specific surgical procedure and potentially introducing bias, either consciously or subconsciously, is not an issue in this type of study design. However, care must still be taken when interpreting the results, ensuring that the RCT methodology is rigorous. The EBRCT design represents an ideal approach to the surgical trial. Presently, this design is not prominent in the literature; however, appears to be growing more frequent in the past decade. Knowledge in how to identify an appropriately designed EBRCT, and analyze and interpret the methodology and results, will no doubt be critical to evidence-based surgery moving forward.

Resolution of Scenario

In follow-up with your patient, you present the results of the Mohtadi et al. [4] paper. You explain that according to the best level of evidence, there is no significant difference between the two treatments in terms of quality of life based on a follow-up of 2 years. Furthermore, you explain that the function of the shoulder is similar between the open and arthroscopic treatments. However, you caution the patient that recurrence of shoulder

instability can occur approximately twice as frequently in patients undergoing arthroscopic repair. Following this discussion, the patient wishes to proceed with an open repair of his right shoulder.

Appendix 1 Complete search strategy performed on March 1st, 2024 (Medline and Embase database example)

1. Joint Instability/
2. Shoulder Dislocation/
3. Glenohumeral instability.mp.
4. Shoulder Injuries/
5. Shoulder instability.mp.
6. Open surgery.mp.
7. Open.mp.
8. Arthroscopy/
9. Arthroscopic Surgery.mp.
10. "Quality of Life"/
11. QoL.mp.
12. Or/1-5
13. Or/6-7
14. Or/8-9
15. Or/10-11
16. 12 and 13 and 14 and 15

References

1. van der Linden W. Pitfalls in randomized surgical trials. *Surgery*. 1980;87(3):258–62.
2. Devereaux PJ, Bhandari M, Clarke M, Montori VM, Cook DJ, Yusuf S, Sackett DL, Cinà CS, Walter SD, Haynes B, Schünemann HJ, Norman GR, Guyatt GH. Need for expertise based randomised controlled trials. *BMJ*. 2005;330(7482):88 Review.
3. Waltho D, Kaur MN, Haynes RB, Farrokhyar F, Thoma A. Users' guide to the surgical literature: how to perform a high-quality literature search. *Can J Surg*. 2015;58(5):349–58.
4. Mohtadi NG, Chan DS, Hollinshead RM, Boorman RS, Hiemstra LA, Lo IK, Hannaford HN, Fredine J, Sasyniuk TM, Paolucci EO. A randomized clinical trial comparing open and arthroscopic stabilization for recurrent traumatic anterior shoulder instability: two-year follow-up with disease-specific quality-of-life outcomes. *J Bone Joint Surg Am*. 2014;96(5):353–60.
5. Oleckno W. *Epidemiology: concepts and methods*. Illinois: Waveland Press, Inc; 2008.

6. Kirkley A, Griffin S, McLintock H, Ng L. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario shoulder instability index (WOSI). *Am J Sports Med.* 1998;26(6):764–72.
7. Kocher MS, Horan MP, Briggs KK, Richardson TR, O'Holleran J, Hawkins RJ. Reliability, validity, and responsiveness of the American shoulder and elbow surgeons subjective shoulder scale in patients with shoulder instability, rotator cuff disease, and glenohumeral arthritis. *J Bone Joint Surg Am.* 2005;87(9):2006–11.
8. Guyatt GH, Rennie D, Meade MO, Cook DJ. *Users' guides to the medical literature: a manual for evidence-based clinical practice.* 3rd ed. New York: McGraw-Hill; 2015.
9. Min K, Fedorka C, Solberg MJ, Shaha SH, Higgins LD. The cost-effectiveness of the arthroscopic Bankart versus open Latarjet in the treatment of primary shoulder instability. *J Shoulder Elb Surg.* 2018;27(6S):S2–9.

Cluster-Randomized Trials in Surgery

15

Kyle Gouveia, Sofia Bzovsky, and Sheila Sprague

Abbreviations

CHG	Chlorhexidine Gluconate
CONSORT	Consolidated Standards of Reporting Trials
DE	Design Effect
EBM	Evidence-Based Medicine
ERAS	Enhanced Recovery After Surgery
GEE	Generalizing Estimating Equations
GLMM	Generalized Linear Mixed Models
ICC	Intra-cluster Correlation Coefficient
IRB	Institutional Review Board
RCT	Randomized Controlled Trial
VTE	Venous Thromboembolism

K. Gouveia
Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
e-mail: kyle.gouveia@medportal.ca

S. Bzovsky
Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada
e-mail: bzovskys@mcmaster.ca

S. Sprague (✉)
Division of Orthopaedic Surgery, Department of Surgery & Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada
e-mail: sprags@mcmaster.ca

Introduction

Randomized controlled trials (RCTs) have become the gold standard for evaluating interventions in the world of evidence-based medicine (EBM) and evidence-based surgery, and the number and quality of surgical RCTs being performed continues to increase over time [1]. Since the introduction of randomization as a means to determine treatment allocation, the standard unit of randomization has been the individual patient [2, 3]. From simple randomization to stratified or block randomization, this is done with the goal of achieving balance between groups of both known and unknown prognostic variables, eliminating potential bias in the allocation of an intervention.

While traditional RCTs with the individual patient as the unit of randomization are effective tools for surgical research, in some circumstances the ability to randomize patients as clusters may be useful. Cluster-randomized trials allow for the randomization of large groups of patients, and specifically are useful for evaluating interventions that operate at a group or population level, or that involve organizational changes [4, 5]. Moreover, the use of cluster randomization is especially effective when attempting to conduct large-scale trials of simple interventions, which are growing in popularity [6]. The cluster unit is quite flexible depending on the needs of the investigators and available study populations,

and for surgical researchers may take the shape of clinics, hospitals, wards, or even individual surgical practices [7, 8].

While the benefits of such a design are apparent, multiple statistical and methodological considerations arise when moving trial design from individual patient randomization to cluster randomization. When randomizing clusters instead of individual patients, the investigator must acknowledge that a limited number of heterogeneous groups makes it very difficult to properly distribute any potential sources of confounding—the initial goal of randomizing treatment allocation. That being said, the ultimate consequence of adopting a cluster design is that the outcomes for individual patients can no longer be assumed to be independent of one another, as with a parallel arm RCT [5]. Consequently, the statistical power of a cluster trial is reduced when compared to a similar sized individual trial. Despite this, cluster trials have the potential to be practice changing and there are multiple circumstances where they should be considered as a possible trial design. This chapter will outline some of the essential points to consider when determining if a research question is best answered by a cluster-randomized trial, important points to consider when designing and undertaking such a trial, as well as limitations. Given the complexity of conducting any large, randomized trial, we will focus on key points relevant specifically to cluster-randomized trials.

When to Consider a Cluster-Randomized Trial

As previously mentioned, cluster-randomized trials carry with them an additional degree of challenge from a statistical perspective, especially as it relates to study power and sample size calculation. However, some research questions in surgery are best answered by this trial design, which invites the question: when should I consider a cluster-randomized trial for my research question? Three factors to consider include the nature of the trial and where it falls

on the pragmatic-explanatory continuum [9, 10], the intervention of interest, and the risk of contamination.

Pragmatic Versus Explanatory

Trials can be thought of as pragmatic or explanatory; however, it is now recognized that most trials exist somewhere on a continuum between the two [9]. While explanatory trials test efficacy of an intervention under ideal conditions—that is, within a strict protocol and with selective inclusion criteria, pragmatic trials investigate the effectiveness of treatment strategies in real world conditions [11, 12]. This results in pragmatic trials aiming to maximize generalizability and often having broad inclusion criteria. Given that these pragmatic trials can focus on interventions taking place on a wider scale, such as institutionally, cluster randomization is often an appropriate choice for evaluating the real-world effectiveness of an intervention. Specifically, when considering randomized trials in surgery, using the example of an intervention in standardized perioperative care—one can see how evaluating this sort of intervention is best done in a pragmatic fashion and that investigators would be interested in determining its effectiveness at the level of the cluster.

The Intervention of Interest

Building from the description of pragmatic trials above, it is essential to understand that ultimately for a research question to best be answered using a cluster-randomized trial, the outcome of interest should be the effectiveness of the intervention at the cluster level. The types of interventions that lend themselves to cluster randomization then are those which evaluate systems level interventions. Such examples in the surgical literature include evaluation of surgical skin antisepsis [13, 14], venous thromboembolism (VTE) prophylaxis [15], and surgical education [16].

Contamination

In keeping with the theme of cluster-randomized trials being useful for evaluating broad, systems level interventions, investigators should understand that trials of these interventions done in the traditional way often increase the risk of contamination within the trial, that is the risk of patients in one group being exposed to the intervention of another group [17]. For example, when applying a broad, systems level intervention such as an enhanced recovery after surgery (ERAS) protocol to surgical patients, there is a risk that individual patients randomized to the control group may still end up being treated at least partially in the same way that patients in the ERAS protocol group are treated. This phenomenon, known as contamination, affects the validity of trial results as it can dilute the observed differences between the intervention and control arms. Randomizing patients at the cluster level can help mitigate this risk, as all patients treated at a certain site or by a certain surgeon will be randomized as a cluster to the same treatment arm [17, 18].

Cluster-Randomized Designs

Parallel Design

The most basic cluster randomized design is the parallel cluster-randomized trial, where clusters are randomized to the intervention or control once, and observations are taken on the members of those clusters with no crossover of clusters to a different study arm during the trial [19]. As mentioned, one of the downsides of cluster-

randomized trials is their relative lack of statistical power when compared to traditional RCTs with patient-level randomization. This is most apparent in a traditional parallel cluster-randomized design. Two alternative cluster-randomized designs that can help minimize this loss of power and efficiency include stepped wedge and cluster-randomized crossover designs, with each incorporating an element of the cluster acting as its own control (Fig. 15.1).

Stepped Wedge Design

In a stepped wedge design, each cluster crosses over only once from the control group to the intervention at differing time points, and the order that clusters cross over is random (Fig. 15.1) [20]. Once this crossover has occurred, the cluster remains in the intervention arm for the remainder of the trial. These designs also typically incorporate a baseline time point where none of the clusters are receiving the intervention. Stepped wedge designs are overall useful in situations where it would be difficult to get an individual cluster to revert back to the control after the implementation of an intervention. This is especially true in cases where it may be unethical to withhold an intervention, such as evaluating the cluster-level effects of an intervention proven to be effective on the individual level. In surgical trials, stepped wedge designs are often employed for perioperative care protocols, such as ERAS protocols, as one can recognize that after implementation of a multifaceted protocol, it would be difficult to fully cross over and revert back to the original standard of care as a control [21, 22]. Generally

Fig. 15.1 Borrowed from Hussey and Hughes: Design and Analysis of Stepped Wedge Cluster-Randomized Trials [20]

Parallel		Crossover			Stepped Wedge					
Cluster	Time	Cluster	Time		Cluster	Time				
	1		1	2		1	2	3	4	5
	1	1	1	1	0	1	0	1	1	1
2	1	2	1	0	0	2	0	0	1	1
3	0	3	0	1	0	3	0	0	0	1
4	0	4	0	1	0	4	0	0	0	0

speaking, stepped wedge designs are used almost exclusively for nontherapeutic interventions such as protocols or quality improvement interventions. Lastly, while stepped wedge designs offer a statistical power advantage over parallel cluster-randomized designs, they are still not as powerful as cluster crossover designs.

Cluster-Randomized Crossover Trials

In a cluster-randomized crossover design, clusters are randomized once to either the intervention or control group, but then after a set period of time they alternate to the other treatment arm. Additionally, this can occur with or without a washout period to minimize contamination (Fig. 15.1 above uses an example without a washout period). In this design, each cluster can truly act as its own control, and this results in cluster crossover trials being the most statistically efficient cluster-randomized design [23]. Many authors recommend that as long as a bidirectional design is possible, the cluster crossover design should be the design of choice [23]. When designing a cluster crossover trial, investigators need to carefully consider the length of the time periods between crossovers, as well as when the interventions are being applied. This is because some outcomes can be different depending on the seasonal variation.

Conducting a Cluster-Randomized Trial

As with any good surgical trial, a good research question comes first. Once a well-defined and impactful research question has been developed, investigators need to ask the question regarding whether their research question is best answered by a cluster-randomized design. Specifically, researchers should ask if the intervention seeks to act at a cluster level. Once the decision is made to conduct a cluster-randomized trial, there are certain considerations that can vary considerably from traditional RCTs.

Cluster Identification and Selection

As described in the CONSORT extension for cluster-randomized trials, a requirement in this trial design is to provide information on the characteristics and flow of clusters throughout the conduct of such a trial [24]. One of the keys to helping investigators accomplish this is the use of a priori cluster selection criteria, similar to the way that patient eligibility criteria are used in a traditional RCT. Furthermore, a cluster flow diagram can show cluster selection and flow over the course of a trial similar to a patient flow diagram (Fig. 15.2).

Prior to cluster selection, investigators must first define the level of their cluster, whether it be a community, hospital, clinic, or individual surgical practice. The choice of cluster often involves deciding on a unit of randomization at a higher or lower level [25]. Generally speaking, it is more statistically powerful to select clusters at the lowest possible level, ensuring the risk for contamination is considered, as this will result in more clusters [26]. As a trade-off, however, randomizing clusters at a higher level often enhances the associated logistics. As an example, randomizing individual surgeons' practices as the cluster unit of choice would increase statistical efficiency, allowing for more units of randomization. In this context, statistical efficiency refers to the statistical power achieved from a given number of observations under a given design. This is because while increasing both number of clusters and cluster size increases power in a cluster-randomized trial, increasing the number of clusters results in a greater increase in power, all other things being equal [27]. On the other hand, randomizing whole surgical departments as the cluster unit would be much larger and result in a lower number of clusters, but would make the logistics of conducting a trial much easier. Not only would it aid in the ability of the team to adhere to the study protocol, but it would also reduce the risk of contamination. These trade-offs need to be considered when selecting the cluster unit you plan to randomize.

Ultimately, this strict screening and enrollment process helps to minimize the risk of needing to withdraw clusters due to inability to meet trial

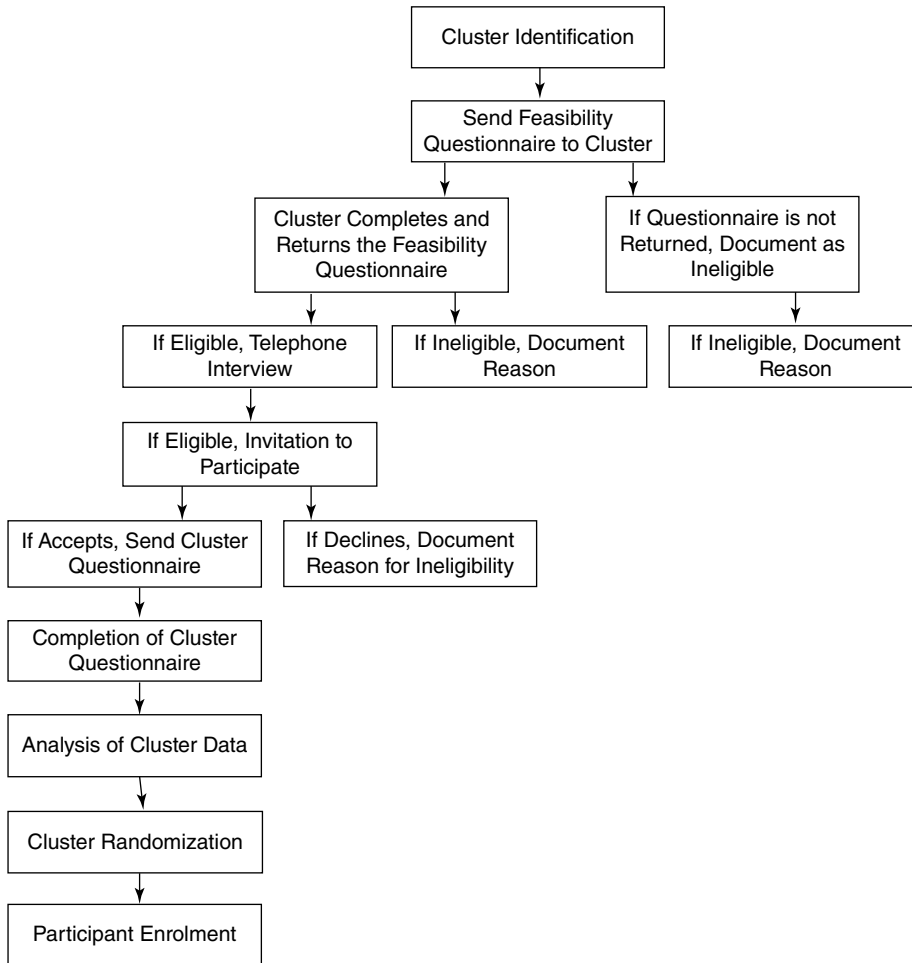


Fig. 15.2 From Sprague et al. 2020: cluster identification, selection, and description in cluster-randomized crossover trials: the PREP-IT trials [8]

protocol. Despite this, unexpected challenges are possible and included in a protocol for a cluster-randomized trial should be a priori reasons for possible cluster withdrawal during conduct of the trial. These typically include the clusters' ability to ensure patients receive the intervention, as well as ensuring proper follow-up and completion of required case report forms (CRFs).

To summarize, proper cluster identification and selection for the clinical trialist falls somewhere between the processes of screening and recruitment as one would for an individual RCT, and the management of sites in the way that a research team would conduct regular evaluation during a multicenter trial.

Pilot Cluster-Randomized Trials

As with any trial design, careful consideration regarding feasibility is necessary prior to going forward with a definitive trial. One such way of ensuring this is by performing a pilot or feasibility trial before continuing on with the definitive trial. Feasibility studies can serve multiple purposes, including both informing the number and sizes of clusters needed for a definitive trial as well as determining the ability of clusters to adhere to study protocol. The latter especially requires defined feasibility outcomes, as should be present in all pilot or feasibility studies [28] (see Chap. 12). For a cluster-randomized study,

such as a cluster crossover trial, a reasonable approach to a feasibility study would be to conduct a pilot study utilizing only a small number of clusters compared to the planned definitive trial [29]. Another potential method of ensuring acceptable compliance among clusters is by utilizing a run-in period prior to initiating participant enrollment [30].

Cluster-randomized trials are different in that clusters are typically higher-level systems, whether it be hospitals or clinics, most cluster-randomized trials need to be treated similarly to multicenter trials. If choosing a cluster-randomized design, investigators should understand that determining feasibility of the study should include determining the ability of the various clusters to adhere to study protocols in the way that investigators of a multicentered trial would ensure various sites can adhere.

Randomization

To minimize the risk of period effects having an impact on the observed effect sizes, blocking randomization periods and creating short crossover periods are often necessary in cluster crossover designs. Despite this, cluster crossover designs are by far still the most statistically efficient cluster-randomized trial design and should be considered where possible.

Reporting in Cluster-Randomized Trials

Initially published in 1996, the Consolidated Standards of Reporting Trials (CONSORT) statement was developed to standardize reporting in RCTs [31]. Since then, it has become the gold standard, with multiple extensions to the statement being published for various trial designs. This includes an extension for cluster-randomized trials which was published in 2012 [24]. While the entire scope of the CONSORT extension for cluster-randomized designs will not be discussed in this chapter, a few key features are worth mentioning.

Firstly, with regard to the reporting of study design, authors should clearly justify the use of cluster randomization, since this unit of randomization taints the idea of prognostic balance that comes with individual randomization. This justification for cluster randomization should also come with a clear definition of the unit of randomization and how statistical power calculations or sample size calculations account for between-cluster variation. With regard to the reporting of results, authors should make clear whether inferences are primarily directed at the cluster level or at the level of the individual subject. Lastly, the reported statistical analyses should account for between-cluster variation and should adjust for prognostically important baseline risk factors.

Challenges and Risks of Bias

Like any study design, cluster-randomized trials come with their unique concerns regarding sources of bias. The primary form of bias to consider is that patients within a cluster are typically more similar to each other than patients between clusters, and therefore are more likely to have similar outcomes. This is measured by the intra-cluster correlation (ICC), which will be discussed further in this chapter. When thinking about a cluster unit, such as an individual hospital, it is easy to understand that one hospital may have certain demographics to its patients, or certain practices within its policies and procedures that can affect outcomes. It is in this way that cluster-randomized trials do not always achieve what is the goal of parallel arm RCTs—balance between groups of prognostic factors both known and unknown. One such way to combat this is the use of a cluster crossover design, where each cluster can effectively act as its own control.

An additional source of bias to consider is that often patients need to be identified or recruited following a cluster's randomization. In that sense, patient allocation cannot be concealed and selection bias can operate in an unpredictable fashion [32]. Investigators will already know to which intervention a patient has been assigned prior to enrollment, introducing the opportunity for

researchers to introduce selection bias. Lastly, even when utilizing a cluster crossover design, in which each cluster acts as its own control, investigators need to be aware that the time periods used for intervention phases can introduce bias, as many outcomes and prognostic factors are affected by seasonality.

Statistical Considerations in Cluster-Randomized Trials

The key characteristic of cluster-randomized trials that needs to be considered in all statistical processes is that the individual participants within a cluster are likely correlated. This affects all statistical assumptions, plays a part in sample size calculations and trial analysis, and supports the validity of the trial results. This section will briefly touch on some key statistical considerations for cluster-randomized trials, as this is a complex and evolving field.

The Intra-Cluster Correlation Coefficient

A commonly seen term in the discussion of cluster-randomized trials is the intra-cluster correlation coefficient, or ICC. This value, which ranges from 0 to 1, is a measure of the relatedness of clustered data, which is relevant as similarity amongst patients in clusters reduces the variability of responses in a clustered sample [33]. This in turn leads to decreased power in the ability to detect true differences between arms. With an ICC of 0, there is no correlation of outcomes within a cluster, and with an ICC of 1, all responses within a cluster are identical. With a higher ICC, the effective sample size is reduced, resulting in a need for more clusters to achieve desired statistical power. This value is important for estimating required sample sizes, and a pilot trial can be helpful in estimating an ICC. However, investigators should understand that while within cluster variance can be well estimated this way, between cluster variance is more difficult to estimate due to a smaller number of clusters [34].

Unit of Inference

While the cluster is the unit of randomization, investigators often must choose whether the cluster or the individual patient will be the unit of inference or the unit of analysis. For many with a focus on surgical trials, the individual patient is more likely to be the chosen unit of inference, rather than the whole cluster in community or population-based studies. Given that we know observations within clusters tend to be correlated, this decreases the effective sample size. Ultimately, the choice of the unit of inference in a cluster-randomized trial should be made to align with the research question, because carefully considering the target of inference is essential for determining appropriate analysis [35].

Sample Size

As previously mentioned, sample sizes in cluster-randomized trials often need to be much larger due to their relative statistical inefficiency. Sample size calculation in cluster-randomized trials requires taking a calculated sample size for an individual randomized trial and utilizing a correction factor known as the design effect (DE) [36]. This value takes into consideration the ICC (given by the Greek rho [ρ]) as well as the average cluster size (n), in the formula $DE = 1 + (n - 1)\rho$. This value represents the degree of variance inflation attributable to cluster sampling [37]. With a calculated design effect, the effective sample size can be calculated by dividing the total number of participants (patients per cluster multiplied by number of clusters) by the design effect. Using the understanding of these formulas, investigators can understand how these factors will affect the needed sample sizes. Holding the number of total patients constant while increasing the number of clusters lowers the design effect, resulting in a greater effective sample size (ESS). Generally speaking, a high number of clusters and a low number of elements per cluster yield the smallest design effect [33]. For the clinical trialist, the major takeaway to remember here is that in study design, increasing

the number of clusters will increase the study's power more than increasing the cluster size.

Analysis

Clustering should always be accounted for in the statistical analysis to support the validity of the results and should be aligned with the unit of inference—that is, analysis should be at the level of the cluster or the individual. Carrying out cluster-level analysis is the simplest approach, as one can aggregate cluster-level outcomes using a summary statistic (such as a mean or proportion), and then use conventional analysis methods such as a t-test [37, 38]. However, these cluster level approaches are infrequently used, as they do not work well when there is substantial variation in cluster size [39]. Additionally, the complexity of the cluster-level approach is much greater when one wants to adjust for individual-level covariates [40]. While the finer details are beyond the scope of this chapter, for those choosing to use individual-level analysis for a cluster-randomized trial, the two main approaches are generalized linear mixed models (GLMM) and generalizing estimating equations (GEE) [41]. Lastly, it should be recognized that the above two approaches do not work well with fewer than 40 to 50 clusters, unless combined with a small sample correction [42]. To summarize, the complexity associated with cluster-randomized trial analysis should always necessitate the involvement of an experienced biostatistician and should involve thoughtful consideration about the unit of analysis, based upon the research question.

Ethical Considerations in Cluster-Randomized Trials

The primary ethical issue raised in cluster randomized trials when compared to traditional RCTs is the issue of consent. Given the nature of the interventions, patients are not always able to choose to avoid or consent to an intervention being performed at the level of a cluster. This is actually one of many circumstances in clinical research where the traditional model of consent

is not always possible, along with patients who are critically ill or injured and require immediate intervention [43]. In these circumstances, ethics committees or institutional review boards (IRBs) may consider waiving some or all of the elements of traditional informed consent, as long as the proposed benefits outweigh risks to participants. Such a waiver of consent may be granted as long as the research involves no more than minimal risk to participants, the waiver or alteration of consent will not adversely affect the rights and welfare of the research participant, and that the research would be impossible or impractical without the waiver or alteration [43, 44].

Given the challenges with informed consent in cluster-randomized trials, a number of these trials have employed the alternative models of deferred consent or waiver of consent, in addition to traditional consent models [45–49].

To summarize, the nature of cluster-randomized trials often requires the investigators to consider an alternative model of consent, and as long as the proposed trial meets requirements laid out by the IRB or ethics committee, these alternative consent models are beneficial to study practicality.

Clinical Scenario

You are an orthopedic surgeon who frequently treats patients with closed fractures. Prior to surgery, you wash the limb with chlorhexidine gluconate (CHG)-based solution in an alcohol. One of your patients who had undergone surgery for a closed tibia fracture and was treated with CHG in alcohol presents to your clinic for a 6-week follow-up appointment with a surgical site infection. You remember that you recently listened to a podcast mentioning a cluster-randomized crossover trial that evaluated preparation solutions in fractures and suggested that iodine povacrylex in an alcohol solution may reduce the risk of infection. You wonder whether you should be using iodine povacrylex in an alcohol solution as preoperative antisepsis for your closed fracture patients and decide to review the literature to answer this question.

Literature Search

You begin your literature search by formulating a research question using the PICOT format: “In patients with a closed fracture requiring surgical fixation, does washing the limb prior to surgery with iodine povacrylex in an alcohol solution reduce the risk of a surgical site infection within 90 days compared to using a CHG-based solution in alcohol?” Using OVID Medline, you enter the following search terms: “preoperative antisepsis AND closed fracture AND cluster-randomized crossover trial AND iodine povacrylex AND chlorhexidine gluconate AND alcohol” and limit your search to manuscripts including humans and published in English in the last 3 years ([Appendix](#)). Your search produces one article titled, “Skin Antisepsis before Surgical Fixation of Extremity Fractures” [14]. The article contains the results of the PREPARE trial, mentioned in the podcast, which followed a cluster-randomized crossover design where investigators randomized orthopaedic practices (clusters) to use either iodine povacrylex in an alcohol solution versus a CHG-based solution in alcohol as preoperative antisepsis on patients undergoing surgery to treat a closed lower extremity or pelvic fracture [14]. This paper appears to address your clinical question and you wonder if you should incorporate their findings into your surgical practice. To evaluate the quality, and the applicability to your practice, you evaluate this cluster trial following the principles of evidence-based surgery. You critique the published paper using the following questions: (a) Is this study valid?, (b) What are the results?, and (c) Can I apply the results of this cluster RCT in my practice? (Box 15.1) [50, 51].

Box 15.1 Important Points to Consider When Evaluating a Cluster RCT

Are the results valid?

- Was full information provided on the number of clusters approached, recruited, and analyzed?

- How was cluster randomization and allocation performed?
- Was clustering accounted for in the sample size calculation?
- Was clustering accounted for in the analysis?

What are the results?

- What impact did the treatment have?
- How precise were the results?

Are the results applicable to my practice?

- How clinically relevant are the results?
- Do the results apply to my patient population?
- How do these results impact me?

This list was modified from Guyatt et al. and Eldridge et al. [50, 51]

Are the Results Valid?

Was Full Information Provided on the Number of Clusters Approached, Recruited, and Analyzed?

Investigators of the PREPARE trial chose orthopedic practices as their cluster level, and laid out defined inclusion and exclusion criteria for their clusters as one would do for individual patients in a traditional RCT [8]. Once this had been done, investigators reached out to orthopedic practices as potential clusters with a feasibility questionnaire, allowing screening for cluster eligibility to take place. After an additional screening, further meetings were used to determine if clusters could successfully function within the trial (taking into account practical aspects, research personnel, etc.) (Fig. 15.2). These procedures were outlined in detail in the primary paper.

How Was Cluster Randomization and Allocation Performed?

The order of treatment allocation for each orthopedic practice (i.e., cluster) was randomly assigned

using a computer-generated randomization table in the PREPARE trial. Simple, non-stratified randomization was used and performed centrally at the Methods Centre by personnel who were unaware of the cluster. Each cluster started with the surgical preparation solution to which they were initially allocated and subsequently crossed over to the other solution for their second cluster period. This process of alternating treatments repeated approximately every 2 months as dictated by the initial randomization. This process of alternating treatment periods (crossovers) continued until the minimum sample size was achieved [8].

Was Clustering Accounted for in the Sample Size Calculation?

Investigators of the PREPARE trial calculated that they would need a sample size of 6280 closed fracture patients to yield 80% or more power to detect differences between the treatment groups. For their calculation, they assumed that compared with CHG, iodine povacrylex in an alcohol solution would achieve a 0.65 risk ratio for SSIs, a between-cluster ICC of 0.028, a between-cluster variance of 0.095, no between-period variance, 10 recruiting clusters, and a 10% loss to follow-up rate [14].

Was Clustering Accounted for in the Analysis?

Investigators of the PREPARE trial evaluated treatment effects on the trial outcomes using mixed-effects regression models with binomial distributions. The closed-fracture model included fixed effects for trial group, a continuous measure of the chronologic recruitment period (time in days), and indicators for severe soft-tissue injuries and periarticular fractures. PREPARE trial investigators also included a cluster indicator as a random intercept with an exchangeable correlation structure and confirmed that the random effects were normally distributed and that the results were robust to specifying time as a categorical indicator or nonlinear term. For the primary outcome, the intraclass correlation was 0.03 [14].

What Are the Results?

What Impact Did the Treatment Have?

How Precise Were the Results?

This trial found that patients with closed fractures randomized to receive iodine povacrylex in an alcohol solution for skin antisepsis experienced fewer postoperative surgical site infections than those randomized to receive CHG in an alcohol-based solution (2.4% versus 3.3%; odds ratio, 0.74; 95% confidence interval, 0.55 to 1.00; $P = 0.049$) [14].

Are the Results Applicable to My Practice?

How Clinically Relevant Are the Results?

Based on the results of this trial, the PREPARE investigators concluded that the use of iodine povacrylex in an alcohol solution as preoperative skin antisepsis could prevent surgical site infection in thousands of patients with closed fractures. Many hospitals may consider a policy change to use iodine povacrylex in an alcohol solution for all fracture surgeries for this reason [14].

Do the Results Apply to My Patient Population?

The PREPARE trial enrolled adults (≥ 18 years old) who were undergoing surgical fixation of a closed lower-extremity or pelvic fracture. As you often treat these types of patients at your hospital, you realize that the PREPARE trial results apply well to your patient population.

How Do These Results Impact Me?

Upon reviewing these evaluation questions for validity and realizing that these are clinically relevant results that apply to your patient population, you decide that you will make the change to use iodine povacrylex in an alcohol solution as the preoperative antisepsis into your surgical practice when treating closed fracture patients in hopes to try to reduce their risk of infection.

Summary

Cluster-randomized trials are statistically less efficient than traditional RCTs, requiring larger sample sizes and they are at a higher risk of bias than a parallel arm RCT. When the intervention is truly delivered at the level of the cluster, cluster randomization is necessary and appropriate. Furthermore, cluster randomization can also help increase the generalizability of findings. Investigators should be thoughtful about the applicability of their research question to a cluster-randomized design, and carefully weigh the benefits of cluster randomization with the increased sample size and the risks of bias it introduces. Ultimately, with the right research question and so long as investigators keep in mind the special methodological and statistical considerations specific to cluster-randomized designs, many successful surgical trials using this design are possible.

Appendix

Database: Ovid MEDLINE(R) ALL < 1946 to September 26, 2024>

Search Strategy:

1. (preoperative antisepsis and closed fracture and cluster-randomized crossover trial and iodine povacrylex and chlorhexidine gluconate and alcohol).mp. [mp = title, book title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms, population supplementary concept word, anatomy supplementary concept word] (1).
2. limit 1 to (english language and humans and yr. = "2021 -Current") (1).

References

1. Pronk AJM, Roelofs A, Flum DR, Bonjer HJ, Abu Hilal M, Dijkgraaf MGW, Besselink MG, Ahmed AU. Two decades of surgical randomized controlled trials: worldwide trends in volume and methodological quality. *Br J Surg*. 2023;110(10):1300–8.
2. Medical Research Council Investigators. Streptomycin treatment of pulmonary tuberculosis. *Br Med J*. 1948;2(4582):769–82.
3. D'Arcy Hart P. A change in scientific approach: from alternation to randomised allocation in clinical trials in the 1940s. *BMJ*. 1999;319(7209):572–3.
4. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health*. 2004;94(3):423–32.
5. Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. The implications of adopting a cluster design are still largely being ignored. *BMJ*. 1998;317(7167):1171–2.
6. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. *J Clin Epidemiol*. 1995;48(1):23–40.
7. Turner EL, Platt AC, Gallis JA, Tetreault K, Easter C, McKenzie JE, Nash S, Forbes AB, Hemming K, CRT Binary Outcome Reporting Group. Completeness of reporting and risks of overstating impact in cluster randomised trials: a systematic review. *Lancet Glob Health*. 2021;9(8):e1163–8.
8. Sprague S, Scott T, Dods S, Pogorzelski D, McKay P, Harris AD, Wood A, Thabane L, Bhandari M, Mehta S, Gaski G, Boulton C, Marciano-Fernández F, Guerra-Farfán E, Hebden J, O'Hara LM, Slobogean GP. PREP-IT Investigators. Cluster identification, selection, and description in cluster randomized crossover trials: the PREP-IT trials. *Trials*. 2020;21(1):712.
9. Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, Tunis S, Bergel E, Harvey I, Magid DJ, Chalkidou K. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol*. 2009;62(5):464–75. <https://doi.org/10.1016/j.jclinepi.2008.12.011>.
10. Zatzick D, Palinkas L, Chambers DA, Whiteside L, Moloney K, Engstrom A, Prater L, Russo J, Wang J, Abu K, Iles-Shih M, Bulger E. Integrating pragmatic and implementation science randomized clinical trial approaches: a PRagmatic Explanatory Continuum Indicator Summary-2 (PRECIS-2) analysis. *Trials*. 2023;24(1):288.
11. Merali Z, Wilson JR. Explanatory versus pragmatic trials: An essential concept in study design and interpretation. *Clin Spine Surg*. 2017;30(9):404–6.
12. Ford I, Norrie J. Pragmatic trials. *N Engl J Med*. 2016;375(5):454–63.

13. PREP-IT Investigators. Aqueous skin antisepsis before surgical fixation of open fractures (Aqueous-PREP): a multiple-period, cluster-randomised, cross-over trial. *Lancet*. 2022;400(10360):1334–44.
14. PREP-IT Investigators, Sprague S, Slobogean G, Wells JL, O'Hara NN, Thabane L, Mullins CD, Harris AD, Wood A, Viskontas D, Apostle KL, O'Toole RV, Joshi M, Johal H, Al-Asiri J, Hymes RA, Gaski GE, Pilson HT, Carroll EA, Babcock S, Halvorson JJ, Romeo NM, Matson CA, Higgins TF, Marchand LS, Bergin PF, Morellato J, Van Demark RE, Potter GD, Gitajn IL, Chang G, Phelps KD, Kempton LB, Karunakar M, Jaeblohn T, Demyanovich HK, Domes CM, Kuhn GR, Reilly RM, Gage MJ, Weaver MJ, von Keudell AG, Heng M, McTague MF, Alnasser A, Mehta S, Donegan DJ, Natoli RM, Szatkowski J, Scott AN, Shannon SF, Jeray KJ, Tanner SL, Marmor MT, Matityahu A, Fowler JT, Pierrie SN, Beltran MJ, Thomson CG, Lin CA, Moon CN, Scolaro JA, Amirhekmata A, Leonard J, Pogorzelski D, Bzovsky S, Heels-Ansdell D, Szasz OP, Gallant JL, Della Rocca GJ, Zura RD, Hebden JN, Patterson JT, Lee C, O'Hara LM, Marvel D, Palmer JE, Friedrich J, D'Alleyrand JCG, Rivera JC, Mossuto F, Schrank GM, Guyatt G, Devereaux PJ, Bhandari M, The PREP-IT Investigators. Skin antisepsis before surgical fixation of extremity fractures. *N Engl J Med*. 2024;390(5):409–20.
15. CRISTAL Study Group, Sidhu VS, Kelly TL, Pratt N, Graves SE, Buchbinder R, Adie S, Cashman K, Ackerman I, Bastiras D, Brighton R, Burns AWR, Chong BH, Clavisi O, Cripps M, Dekkers M, de Steiger R, Dixon M, Ellis A, Griffith EC, Hale D, Hansen A, Harris A, Hau R, Horsley M, James D, Khorshid O, Kuo L, Lewis P, Lieu D, Lorimer M, MacDessi S, McCombe P, McDougall C, Mulford J, Naylor JM, Page RS, Radovanovic J, Solomon M, Sorial R, Summersell P, Tran P, Walter WL, Webb S, Wilson C, Wysocki D, Harris IA. Effect of aspirin vs enoxaparin on symptomatic venous thromboembolism in patients undergoing hip or knee arthroplasty: The CRISTAL Randomized Trial. *JAMA*. 2022;328(8):719–27.
16. Bilimoria KY, Chung JW, Hedges LV, Dahlke AR, Love R, Cohen ME, Hoyt DB, Yang AD, Tarpley JL, Mellinger JD, Mahvi DM, Kelz RR, Ko CY, Odell DD, Stulberg JJ, Lewis FR. National Cluster-Randomized Trial of duty-hour flexibility in surgical training. *N Engl J Med*. 2016;374(8):713–27.
17. Pence BW, Gaynes BN, Thielman NM, Heine A, Mugavero MJ, Turner EL, Quinlivan EB. Balancing contamination and referral bias in a randomized clinical trial: An application of pseudo-cluster randomization. *Am J Epidemiol*. 2015;182(12):1039–46.
18. McCann RS, van den Berg H, Takken W, Chetwynd AG, Giorgi E, Terlouw DJ, Diggle PJ. Reducing contamination risk in cluster-randomized infectious disease-intervention trials. *Int J Epidemiol*. 2018;47(6):2015–24.
19. National Institutes of Health. Parallel Group- or Cluster-Randomized Trials | Research Methods Resources. July 2024. <https://researchmethodsresources.nih.gov/methods/grt>. Accessed 22 Aug 2024.
20. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials*. 2007;28(2):182–91.
21. Dou D, Jia Y, Yuan S, Wang Y, Li Y, Wang H, Ding J, Wu X, Bie D, Liu Q, An R, Yan H, Yan F. The protocol of Enhanced Recovery After Cardiac Surgery (ERACS) in congenital heart disease: a stepped wedge cluster randomized trial. *BMC Pediatr*. 2024;24(1):22.
22. Piovano E, Puppo A, Camanni M, Castiglione A, Delpiano EM, Giacometti L, Rolfo M, Rizzo A, Zola P, Ciccone G, Pagano E, ERAS-Gyneco Piemonte Group. Implementing Enhanced Recovery After Surgery for hysterectomy in a hospital network with audit and feedback: a stepped-wedge cluster randomized trial. *BJOG*. 2024;131(9):1207–17.
23. Hooper R, Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ*. 2015;350:h2925.
24. Campbell MK, Piaggio G, Elbourne DR, Altman DG, CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. *BMJ*. 2012;345:e5661.
25. Murray D, Taljaard M, Turner E, George S. Essential ingredients and innovations in the design and analysis of group-randomized trials. *Annu Rev Public Health*. 2020;41. <https://pubmed.ncbi.nlm.nih.gov/31869281/>. Accessed 22 Aug 2024.
26. Hemming K, Taljaard M. Key considerations for designing, conducting and analysing a cluster randomized trial. *Int J Epidemiol*. 2023;52(5):1648–58.
27. Hemming K, Eldridge S, Forbes G, Weijer C, Taljaard M. How to design efficient cluster randomised trials. *BMJ*. 2017;358:j3064.
28. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, Lancaster GA, PAFS consensus group. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ*. 2016;355:i5239.
29. Sprague S, Guyatt P, Bzovsky S, Nguyen U, Bhandari M, Thabane L, Petrisor B, Johal HS, Leonard J, Dodds S, Mossuto F, O'Toole RV, Howe A, Demyanovich HK, Camara M, O'Hara NN, Slobogean GP, PREP-IT Investigators. Pragmatic randomized trial evaluating pre-operative aqueous antiseptic skin solution in open fractures (Aqueous-PREP): the feasibility of a cluster randomized crossover study. *Pilot Feasibil Stud*. 2021;7(1):61.
30. Program of Randomized Trials to Evaluate Pre-operative Antiseptic Skin Solutions in Orthopaedic Trauma (PREP-IT) Investigators, Slobogean GP, Sprague S, Wells J, Bhandari M, Rojas A, Garibaldi A, Wood A, Howe A, Harris AD, Petrisor BA, Mullins DC, Pogorzelski D, Marvel D, Heels-Ansdell D, Mossuto F, Grissom F, Del Fabbro G, Guyatt GH, Della Rocca GJ, Demyanovich HK, Gitajn IL, Palmer

- J, D'Alleyrand JC, Friedrich J, Rivera J, Hebden J, Rudnicki J, Fowler J, Jeray KJ, Thabane L, Marchand L, O'Hara LM, Joshi MG, Talbot M, Camara M, Szasz OP, O'Hara NN, McKay P, Devereaux PJ, O'Toole RV, Zura R, Morshed S, Dodds S, Li S, Tanner SL, Scott T, Nguyen U. Effectiveness of iodophor vs chlorhexidine solutions for surgical site infections and unplanned reoperations for patients who underwent fracture repair: the PREP-IT master protocol. *JAMA Netw Open*. 2020;3(4):e202215.
31. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA*. 1996;276(8):637–9.
 32. Easter C, Thompson JA, Eldridge S, Taljaard M, Hemming K. Cluster randomized trials of individual-level interventions were at high risk of bias. *J Clin Epidemiol*. 2021;138:49–59.
 33. Killip S, Mahfoud Z, Pearce K. What is an intra-cluster correlation coefficient? Crucial concepts for primary care researchers. *Ann Fam Med*. 2004;2(3):204–8.
 34. Eldridge SM, Costelloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomised trial be? *Stat Methods Med Res*. 2016;25(3):1039–56.
 35. Jin M, Liu G. Estimand framework: delineating what to be estimated with clinical questions of interest in clinical trials. *Contemp Clin Trials*. 2020;96:106093.
 36. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol*. 2006;35(5):1292–300.
 37. Lakshman RR, Sharp SJ, Ong KK, Forouhi NG. A novel school-based intervention to improve nutrition knowledge in children: cluster randomised controlled trial. *BMC Public Health*. 2010;10:123.
 38. Rao JN, Scott AJ. A simple method for the analysis of clustered of clustered binary data. *Biometrics*. 1992;48(2):577–85.
 39. Ukoumunne OC, Forbes AB, Carlin JB, Gulliford MC. Comparison of the risk difference, risk ratio and odds ratio scales for quantifying the unadjusted intervention effect in cluster randomized trials. *Stat Med*. 2008;27(25):5143–55.
 40. Hayes R, Moulton L. Cluster randomised trials. 2nd ed. Chapman and Hall/CRC; 2017.
 41. Turner EL, Prague M, Gallis JA, Li F, Murray DM. Review of recent methodological developments in group-randomized trials: part 2-analysis. *Am J Public Health*. 2017;107(7):1078–86.
 42. Leyrat C, Morgan KE, Leurent B, Kahan BC. Cluster randomized trials with a small number of clusters: which analyses should be used? *Int J Epidemiol*. 2018;47(1):321–31.
 43. Tri-Council Policy Statement. TCPS 2 Chapter 3: The Consent Process. 2018. https://ethics.gc.ca/eng/tcps2--eptc2_2018_chapter3-chapitre3.html#a
 44. US Food and Drug Administration. U.S. FDA 21 Code of Federal Regulations Title 21. Published online 2019. <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=50.20>
 45. Milstone AM, Elward A, Song X, Zerr DM, Orscheln R, Speck K, Obeng D, Reich NG, Coffin SE, Perl TM, Pediatric SCRUB Trial Study Group. Daily chlorhexidine bathing to reduce bacteraemia in critically ill children: a multicentre, cluster-randomised, crossover trial. *Lancet*. 2013;381(9872):1099–106.
 46. McIntyre LA, Fergusson DA, McArdle T, Fox-Robichaud A, English SW, Martin C, Marshall J, Cook DJ, Graham ID, Hawken S, McCartney C, Menon K, Saginur R, Seely A, Stiell I, Thavorn K, Weijer C, Iyengar A, Muscedere J, Forster AJ, Taljaard M, Canadian Critical Care Trials Group. FLUID trial: a hospital-wide open-label cluster cross-over pragmatic comparative effectiveness randomised pilot trial comparing normal saline to Ringer's lactate. *BMJ Open*. 2023;13(2):e067142.
 47. Rosa RG, Falavigna M, Robinson CC, da Silva DB, Kochhann R, de Moura RM, Santos MMS, Sganzerla D, Giordani NE, Eugênio C, Ribeiro T, Cavalcanti AB, Bozza F, Azevedo LCP, Machado FR, Salluh JIF, Pellegrini JAS, Moraes RB, Hocheegger T, Amaral A, Teles JMM, da Luz LG, Barbosa MG, Birriel DC, Ferraz I de L, Nobre V, Valentim HM, Corrêa E Castro L, Duarte PAD, Tregnago R, Barilli SLS, Brandão N, Giannini A, Teixeira C, ICU Visits Study Group Investigators and the BRICNet. Study protocol to assess the effectiveness and safety of a flexible family visitation model for delirium prevention in adult intensive care units: a cluster-randomised, crossover trial (The ICU Visits Study). *BMJ Open*. 2018;8(4):e021193.
 48. Anderson CS, Arima H, Lavados P, Billot L, Hackett ML, Olavarria VV, Muñoz Venturelli P, Brunser A, Peng B, Cui L, Song L, Rogers K, Middleton S, Lim JY, Forshaw D, Lightbody CE, Woodward M, Pontes-Neto O, De Silva HA, Lin RT, Lee TH, Pandian JD, Mead GE, Robinson T, Watkins C, HeadPoST Investigators and Coordinators. Cluster-randomized, crossover trial of head positioning in acute stroke. *N Engl J Med*. 2017;376(25):2437–47.
 49. van Oostveen RB, Romero-Palacios A, Whitlock R, Lee SF, Connolly S, Carignan A, Mazer CD, Loeb M, Mertz D. Prevention of Infections in Cardiac Surgery study (PICS): study protocol for a pragmatic cluster-randomized factorial crossover pilot trial. *Trials*. 2018;19(1):688.
 50. Guyatt G, Rennie D, Meade M, Cook D, American Medical Association. Users' guides to the medical literature: a manual for evidence-based clinical practice. McGraw-Hill Education; 2015.
 51. Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. Internal and external validity of cluster randomised trials: systematic review of recent trials. *BMJ*. 2008;336(7649):876–80. <https://doi.org/10.1136/bmj.39517.495764.25>.

The Surgeon's Guide to Systematic Review and Meta-Analysis

16

Lucas Gallo, Patrick Kim, Andrea Copeland,
Noor Alolabi, and Achilles Thoma

Abbreviations

AMSTAR	A MeaSurement Tool to Assess systematic Reviews
CI	Confidence Interval
CMR	Capture Mark Recapture
COMET	Core Outcome Measures in Effectiveness Trials
COS	Core Outcome Sets
LOE	Level of Evidence
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analysis
PROSPERO	International Prospective Register of Systematic Reviews
RCT	Randomized Controlled Trial

Introduction

As the body of surgical literature expands, the need to summarize research in the form of high-quality reviews and meta-analyses increases. Review articles represent an efficient way to digest and summarize the body of available literature to facilitate up-to-date, evidence-based clinical practice [1–3].

Many types of review articles exist, notably traditional narrative reviews, systematic reviews, and meta-analyses. However, these vary regarding their methodological rigor, their susceptibility to bias, and the accuracy of their results (Table 16.1) [1].

L. Gallo (✉)
Division of Plastic and Reconstructive Surgery,
Department of Surgery, McMaster University,
Hamilton, ON, Canada
e-mail: lucas.gallo@medportal.ca

P. Kim
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: patrick.kim1@medportal.ca

A. Copeland
Breast Reconstruction and Microsurgery Fellow,
Royal Brisbane and Women's Hospital,
Brisbane, QLD, Australia
e-mail: andrea.copeland@medportal.ca

N. Alolabi
Division of Plastic and Reconstructive Surgery,
Oakville Trafalgar Memorial Hospital,
Oakville, ON, Canada

Faculty of Health Sciences McMaster University,
Hamilton, ON, Canada
e-mail: noor.alolabi@medportal.ca

A. Thoma
Division of Plastic Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: athoma@mcmaster.ca

Table 16.1 Differences between traditional narrative reviews, systematic reviews, and meta-analyses^a [1, 4–7]

	Narrative review	Systematic review	Meta-analysis
<i>Research question</i>	Broad	Focused	Focused
<i>Literature search</i>	Not reported, or not comprehensive	Explicit, reproducible	Explicit, reproducible
<i>Article selection criteria</i>	Not reported	Criterion-based, complete	Criterion-based, complete
<i>Quality assessment</i>	Not reported	Methodology assessed	Methodology assessed
<i>Result summary</i>	Qualitative	Qualitative	Quantitative

This table highlights the key differences between narrative reviews, systematic reviews and meta-analyses.

While traditional narrative reviews provide a general overview of a particular topic or question [1], systematic reviews attempt to address a focused clinical question and utilise an explicit and reproducible strategy to identify, appraise, and summarize the primary literature [3]. In meta-analyses, quantitative results are then synthesized to produce a single best estimate of effect. Systematic reviews and meta-analyses are typically restricted to a specific population, intervention, comparator, and outcome—for example, the use of nailing vs. plating in the management of humeral neck fractures in trauma patients and its effect on postoperative complication rates at specific times [1]. Since the first edition of this book, a variant of the meta-analysis design, called the network meta-analysis, has become more common in the literature. A network meta-analysis can compare three or more different interventions to each other [8]. The characteristics of this study design are discussed more thoroughly in Chap. 32 for interested readers.

Systematic reviews and meta-analyses are generally at the top of the level of evidence (LOE) hierarchy but still have limitations [6, 7, 9]. Their quality is dependent on the quality of the primary studies being reviewed, and therefore their value, as a summary of the primary literature, is limited by the input: colloquially this is summarized as “garbage in, garbage out” [10].

This chapter seeks to familiarize surgeons with the basic techniques for critically appraising systematic reviews and meta-analyses as presented in the surgical literature.

Clinical Scenario

You are a plastic surgeon and are asked to see a 55-year-old woman who presents to your office for consultation and consideration for immediate

breast reconstruction after mastectomy. She is scheduled for mastectomy in 2 weeks. After completing your history and physical examination, you believe that she would be a good candidate for implant-based reconstruction as she will be having bilateral skin-sparing mastectomies. As she has minimal abdominal tissue, she is precluded from autogenous abdominal tissue reconstruction. You explain your recommendation to her, and she is agreeable to your plan. However, the patient expresses that she has a history of capsular contracture after a breast augmentation 30 years ago and is worried about developing capsular contracture again. She asks you if there are any ways to minimize this risk. You are uncertain if the prepectoral implant placement has any significant impact on the incidence of capsular contracture. You decide to arrange a follow-up to discuss the available evidence.

Finding the Evidence

This scenario can be structured as a clinical research question in PICOT format as in Table 16.2. (see Chap. 3).

This can then be summarized into the following research question: “Among adult patients undergoing immediate implant-based breast reconstruction, does prepectoral implant placement decrease capsular contracture compared to subpectoral implant placement?”

To find the answer to this question, a high-quality systematic review and meta-analysis or randomized controlled trial are preferable, as they lie at the top of the hierarchy of scientific evidence [11].

To find the evidence, you perform a literature search (see Chap. 4). You limit your search to the following inclusion criteria: (1) English language

Table 16.2 Clinical question demonstrated in PICOT format

Population	Patients undergoing immediate implant-based breast reconstruction
Intervention	Prepectoral breast reconstruction
Comparison	Subpectoral breast reconstruction
Outcome	Capsular contracture
Time horizon	Any time after surgery

articles, (2) systematic reviews or RCTs published within the last 10 years, (3) adult (age ≥ 18 years) population, and (4) articles that assess complication rates following prepectoral vs. subpectoral breast reconstruction. You exclude the following: (1) non-randomized observational studies and (2) studies that evaluate autologous breast reconstruction.

Searching the MEDLINE database from the National Library of Medicine, you derive your keywords from the clinical question. You enter the terms “breast reconstruction” AND “implant” AND “prepectoral” AND “subpectoral” AND “complications”; which yields 120 results. You limit the search to “Systematic Review,” narrowing it to 12 articles—Appendix A. There are three articles [12–14], which appear to answer your research question. After reviewing the three articles, you decide to go with the article by Kim et al. titled “Prevention of Postoperative Complications by Prepectoral versus Subpectoral Breast Reconstruction: A Systematic Review and Meta-Analysis” [14]. This article was published in a journal with higher impact factor, has the most recent publication date, and included the greatest number of articles in its review. Thus, you decide to conduct a critical appraisal. You use the following guide, adapted from Bhandari et al. (2004) [1], to help guide your analysis (Table 16.3).

Are the Results Valid?

Did the Review Address a Sensible Clinical Question?

A good research question is both focused and clinically relevant [15]. A focused question yields a consistent treatment effect across the range of Population, Intervention, Comparison, Outcomes, and Time horizons (the “PICOT elements”).

Table 16.3 Users’ guide for how to review systematic reviews

<i>Are the results valid?</i>
Did the review address a sensible clinical question?
Was the search for relevant studies detailed and exhaustive?
Was publication bias assessed?
Was the systematic review of high methodological quality?
Were the primary studies of high methodological quality?
Were assessments of studies reproducible?
<i>What are the results?</i>
Were the results similar from study to study?
What are the overall results of the review?
How precise were the results?
<i>Can I apply the results to patient care?</i>
How can I best interpret the results and apply them to my patients?
Were all patient-important outcomes considered?
Is the benefit worth the potential costs and risks?

Adapted from Bhandari et al. (2004) [1]

Consider the hypothetical example of a systematic review that pools the results from all therapies, both operative and nonoperative, for all types of upper extremity fractures, to generate a single estimate of the effect on union rates. Clinicians would universally agree that this question is too broad, and the results would not be useful [6]. Conversely, asking a question with very narrow inclusion and exclusion criteria may produce too few primary studies to review, and decreases the generalizability of the results [2].

A clinically relevant question, as described by Thoma et al. [16] in their article “Forming the Research Question,” meets at least one of the following four criteria:

1. The intervention is novel.
2. The intervention consumes large healthcare resources.
3. There is a controversy on the effectiveness of the novel procedure as compared with the existing procedure.
4. There is a large cost difference between two prevailing interventions.

For a complete description on forming a surgical research question, please refer to Chap. 3.

The Kim et al. [14] study addresses an appropriately focused clinical question with a well-defined population (patients who underwent prophylactic or therapeutic mastectomy), intervention (prepectoral implant-based reconstruction), and comparator (subpectoral implant-based reconstruction). The primary outcome can be inferred to be postoperative complications based on the title but was not explicitly defined. Other outcomes included patient-reported outcomes, postoperative pain, and upper extremity function. The answer to this clinical question appears to be uncertain in the current body of literature and is therefore clinically relevant. You feel that Kim et al. [14] addresses a sensible clinical question, as per the criteria listed above [16].

Was the Search for Relevant Studies Detailed and Exhaustive?

The ultimate goal of a search strategy for a systematic review is to include the entire body of evidence that is relevant for a given research question [17]. The published and unpublished literature should be searched, including English and non-English journals [2]. The inclusion of journals in multiple languages reduces the risk of language bias. This is because studies with positive results are preferentially accepted to English-language journals, resulting in an overestimation of treatment effect in the systematic review. It also strengthens the generalizability of the results [3, 7].

Most investigators will search common bibliographic databases such as MEDLINE, EMBASE, and Cochrane Controlled Trials Register [see article by Haines et al. (2008) [2] for descriptions and uses of these databases]. However, there is controversy regarding the inclusion of unpublished literature in systematic reviews. Since unpublished studies can be more likely to be methodologically weak, their inclusion may compromise the validity of the meta-analysis. Conversely, omitting them may lead to publication bias (see the following section). Therefore, it is preferable to include unpublished studies in the search strategy and subsequently perform subgroup analysis to determine whether publication type impacts the results. Unpublished

studies can be identified by hand searching conference proceedings and contacting experts in the field [2, 6, 18]. Consulting a medical librarian helps reassure the reader that the search was exhaustive [2, 3]. Publishing the search strategy also ensures transparency and reproducibility. For more details, see Chap. 4.

Moreover, investigators may also choose to evaluate the completeness of a literature search using capture-mark-recapture (CMR) techniques to estimate the population of articles available for a given topic, known as a horizon estimation [19]. While the specifics of CMR methodology may extend beyond the scope of this chapter, it requires that study authors establish a priori search stopping criteria (% of total article estimate). Then, authors should perform a search of likely databases, complete screening to final inclusion, calculate the horizon estimate using CMR techniques and subsequently compare article retrieval with the established horizon estimate to determine whether the initial search stopping criteria is satisfied [19]. This technique encourages authors to continue to search additional sources, and therefore guide efficient search strategies [19].

Finally, two or more independent investigators should be involved in article selection and data abstraction to minimize errors and limit selection bias. In a recent meta-analysis of systematic reviews, 48.4% of systematic reviews used two or more independent data extractors [7].

Kim et al. [14] did not have a defined time frame but included studies up to April 2021. The authors identified their primary databases (PubMed, EMBASE, Cochrane central register of controlled trials), listed their search terms and complete search strategy (Appendix B), making it adequately reproducible. Additionally, Kim et al. [14] screened all articles using two independent reviewers, ultimately limiting the potential for errors, and minimizing selection bias. However, you notice that the authors did not mention if their literature search was performed with the help of a medical librarian. Furthermore, non-English articles and the unpublished literature were excluded and there is no estimate of search completeness using CMR techniques.

Was Publication Bias Assessed?

Publication bias is “the selective publication of research findings based on the magnitude, direction, or statistical significance of the study result” [20]. Small studies with negative results are less frequently submitted or accepted for publication. In a systematic review that excludes the unpublished literature, publication bias may therefore lead to an overestimation of treatment effect and potentially a false-positive result [2, 18, 21, 22].

Publication bias is frequently explored with a funnel plot. This diagram plots the effect size (x-axis) against the sample size (y-axis) of individual studies (Fig. 16.1). In studies with larger sample sizes, shown at the top of the plot, the scattered dots are closer together representing higher precision. Smaller studies, at the bottom of the plot, are located more peripherally, representing over- or underestimation of a treatment effect. If there is no publication bias, the dots create a symmetrical, inverted funnel shape. If, however, there is a paucity of small studies with

negative findings, the plot will appear asymmetrical at the bottom, implying publication bias [23]. To complement this qualitative representation, a statistical test (e.g., rank correlation test) should be utilized, though it has limited power where the meta-analysis is small (<25 studies) [4, 18, 22].

In the study by Kim et al. [14], publication bias was assessed by using Funnel plots (Fig. 16.1) and the Egger linear regression test. You note that no publication bias was noted for any of the quantitative analyses, although Funnel plots and Egger test values were only reported for certain outcomes (capsular contracture rate, capsular contraction rate for postmastectomy radiation subgroup, infection rate).

Was the Systematic Review of High Methodological Quality?

As systematic reviews and meta-analyses become more commonplace in the literature, there have been many tools developed to improve the quality of these articles. Such tools take the

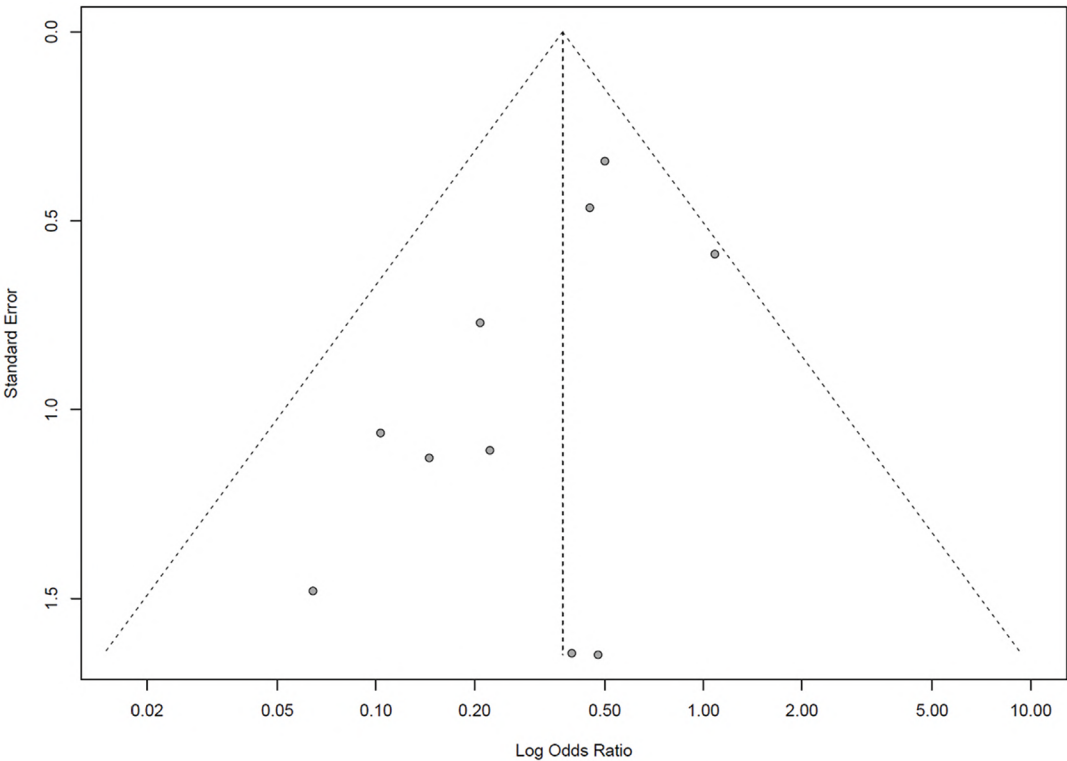


Fig. 16.1 Funnel plot for the comparison of capsular contracture after excluding one study. (Adapted from Kim et al. 2016 [14])

form of databases for protocol registration, guidelines for transparent reporting, and quality assessment tools.

To avoid duplication and promote greater transparency, the International Prospective Register of Systematic Reviews (PROSPERO) database was developed in 2011 for authors of systematic reviews and meta-analyses to prospectively register research protocols where there is at least one health-related outcome [24]. It is recommended that authors publish their protocols as it allows readers to compare it to its completed review and assess for any discrepancies, ultimately minimizing the opportunity for reporting bias [24]. You note that there is no mention of protocol registration in the article by Kim et al. [14].

The Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement [25] is a 27-item checklist and flow diagram that was originally published in 2009 to help authors improve the reporting quality of their systematic reviews and meta-analyses. While not considered to be a quality assessment instrument, the PRISMA statement can be referenced to critically appraise academic review articles published within the literature [24]. In the chosen article by Kim et al. [14], there is no reference to the PRISMA statement and there is no checklist available as a supplementary file.

A MeaSurement Tool to Assess systematic Reviews (AMSTAR) was an 11-item instrument developed in 2007 designed to appraise systematic reviews of RCTs. The AMSTAR has since been updated to the 16-item AMSTAR 2, which can appraise systematic reviews of randomized and non-randomized studies [26]. The 16 items are divided into critical and non-critical domains. Adherence to these domains determine the confidence in the results of the review to be rated between high to critically low, although the authors advise using context and judgement to determine which domains are critical [26]. Based on your assessment of the article by Kim et al. [14], you determine that you have low confidence in the results of the article.

Were the Primary Studies of High Methodological Quality?

The evidence in a systematic review is only as sound as that of its primary articles, thus the methodological quality of primary articles should be assessed. In one study, 35.3% of systematic reviews published in the plastic surgery literature evaluated the methodological quality of the primary studies, and 30% factored this into their conclusions [7].

At present, there is no single, universally endorsed assessment instrument to evaluate the methodological quality of primary studies referenced within systematic reviews and meta-analyses [6]. While several composite scoring tools exist (i.e., checklists and scales), the five-item Jadad score [27], which can be used for the assessment of RCTs (i.e. randomization, double-blinding and participant withdrawal), and the 12-item MINORs score [28], which can be used for the evaluation of nonrandomized trials, are common examples.

While composite scoring tools are commended for their ease of use, it should be noted that many simply record whether methodological safeguards (i.e., allocation concealment) were reported, while others combine their assessment into a composite score. It is more important, however, to assess whether these methodological safeguards were performed appropriately within a trial and present the assessment of each safeguard individually [29]. Domain-based evaluations such as the Cochrane Collaboration's Risk of Bias tool [29], assesses and reports each quality component individually. This "component approach" gives a more transparent analysis of methodological quality and should be utilized [8, 22, 29–32].

Instead of excluding primary studies that are found to have poor methodological quality, investigators should perform a subgroup analysis to determine the impact of the quality score on the results [1].

In the chosen article by Kim et al. [14], methodological quality of the primary studies was appropriately assessed. The MINORs score [28] was used for non-randomized studies, and there were no RCTs identified in the search. Overall, the authors concluded that most included studies

had a low risk of bias, while the remainder had a moderate risk of bias. No studies had a high risk of bias [14]. The authors also provided a breakdown of their risk of bias evaluation for each study, which is more transparent and useful than using scales that simply produce a summative score [6, 29]. However, you notice that they did not use the quality scores to include or exclude any of the studies, nor did they perform subgroup analysis to compare the effect of methodological quality on treatment outcomes.

Were Assessments of Studies Reproducible?

Systematic reviews must have a reproducible selection process. Eligibility criteria and data to be extracted should be listed. The selection process can be visually represented in the form of a PRISMA flow diagram [33]. Each stage should be conducted by two or more independent reviewers to minimize error and bias. Interobserver agreement, such as the kappa statistic, should then be measured. If there is good chance-correlated agreement between investigators, the reader will have more confidence in the results of the review [6, 15]. In addition to the kappa statistic, it is useful to report the variables on which the reviewers most often disagreed. This adds another level of transparency to the selection and data extraction process [23].

The process of study selection by Kim et al. [14] was appropriately documented in a PRISMA flow diagram (Fig. 16.2). However, both study selection and risk of bias assessments were conducted independently by two authors. Additionally, the degree of interobserver agreement (i.e., the kappa statistic) was not reported.

What Are the Results?

Were the Results Similar from Study to Study?

A meta-analysis creates a single summary estimate of treatment effect by combining individual studies as though their sample sizes were all part of one larger study. However, the effect size of

each study should be similar enough to justify combining them. Some variation is expected between studies due to chance. The question becomes, then, to what extent are the observed differences in effect size greater than one would expect by chance.

This question can be answered both visually, by inspecting the point estimates and confidence intervals on a Forest plot, and statistically, by using tests of heterogeneity. In the context of systematic reviews and meta-analyses, heterogeneity refers to the differences in the effect sizes appreciated (intervention outcomes) across included studies. Specifically, similar point estimates and greatly overlapping confidence intervals indicate less heterogeneity and make the pooled summary estimate more meaningful [6, 18].

While visual inspection of the forest plot gives the reader a rough estimate of heterogeneity, statistical tests like the I-squared statistic (I^2) are important to quantify heterogeneity. I^2 (I-squared) describes the percentage of total variation attributable to underlying differences in effect (i.e., heterogeneity) rather than chance [34]. A value above 50% represents substantial heterogeneity. Another test assessing heterogeneity is Cochran's Q test (measured with a chi square test). A low p value (< 0.1) indicates that the observed differences are unlikely to be due to chance. Unfortunately, these tests may lack power to detect statistically significant heterogeneity when the review includes too few studies with few patients in each study [2, 22, 23].

In summary, widely different point estimates, nonoverlapping confidence intervals, a high I-squared (I^2) value, and a low p value associated with the chi-square test should raise concern about pooling results across studies into a single summary estimate.

If substantial heterogeneity exists between studies, authors should look for explanations by performing a subgroup analysis [18], since combining two differing subgroups could lead to misleading conclusions [2, 3]. Other potential sources of heterogeneity may include differences in study groups, interventions, controls, outcome reporting, time horizons, and research

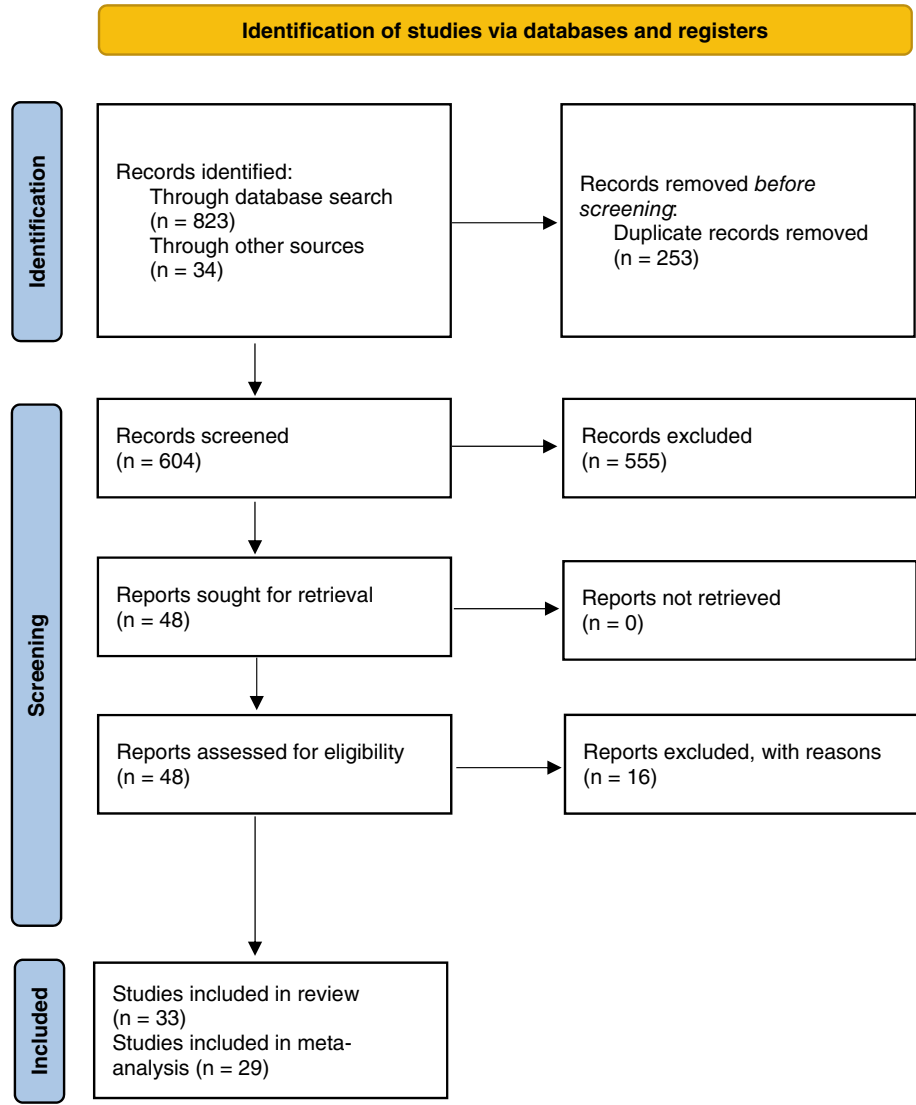


Fig. 16.2 PRISMA flow diagram for study inclusion. (Adapted from Kim et al. 2016 [14])

methodology. These hypotheses of potential sources of heterogeneity should be made a priori, to reduce the likelihood of false positive findings.

Finally, the degree of heterogeneity between studies should direct the choice of statistical effect estimate model. While random-effects models account for variation both within studies and between studies, fixed-effects models assume there is one true effect size underlying all studies and hence do not consider between-study variation. They are therefore only appro-

priate for meta-analyses with a large number of homogenous studies (typically $I^2 < 25\%$). If any degree of heterogeneity is suspected, a random-effects model, while it will produce larger confidence intervals, is the more conservative choice. The reader should be suspicious of a fixed-effects model in the surgical literature, given the inevitable heterogeneity demonstrated by surgery [2, 18].

Kim et al. [14] applied fixed-effects and random-effects models inappropriately based on heterogeneity for each outcome. Heterogeneity

was evaluated using a chi-square-based Q test and I^2 statistic. A fixed-effects model was used when the chi-square test had $p \geq 0.05$ and $I^2 \leq 50\%$, instead of the cutoff of $I^2 < 25\%$. Additionally, as mentioned above we should be wary of the use of a fixed-effects model in a meta-analysis of surgical literature. Instead, Kim et al. [14] opted to use a sensitivity analysis to address heterogeneity unless heterogeneity could not be eliminated, in which case a random-effects model was used.

Although Kim et al. [14] did not designate a primary outcome, we can examine our outcome of interest (capsular contracture rate). The chi-square test and I^2 statistics demonstrated significant heterogeneity ($p = 0.02$, $I^2 = 52.37\%$). Sensitivity analysis excluded one study from the analysis to eliminate heterogeneity ($p = 0.476$, $I^2 = 0\%$), but the authors did not provide any hypotheses as to why this study may have been heterogeneous.

What Are the Overall Results of the Review?

After completing the systematic review, if the primary studies are of poor quality or are too heterogeneous, the data should be presented qualitatively. However, a meta-analysis is performed if the data are appropriate for pooling. The rationale for pooling should be also stated.

Not all studies need to be pooled. For example, nonrandomized data could be included in the systematic review but excluded from the meta-analysis. Alternatively, all studies could be included in the meta-analysis, but the RCTs could be analyzed separately from the non-RCTs [3].

Results from a meta-analysis are typically displayed graphically as Forest plots (Fig. 16.3). Outcomes for dichotomous variables are expressed as ratios, while continuous outcome measures are expressed as weighted or standard mean differences. Individual study results are presented in rows. A box represents the study's

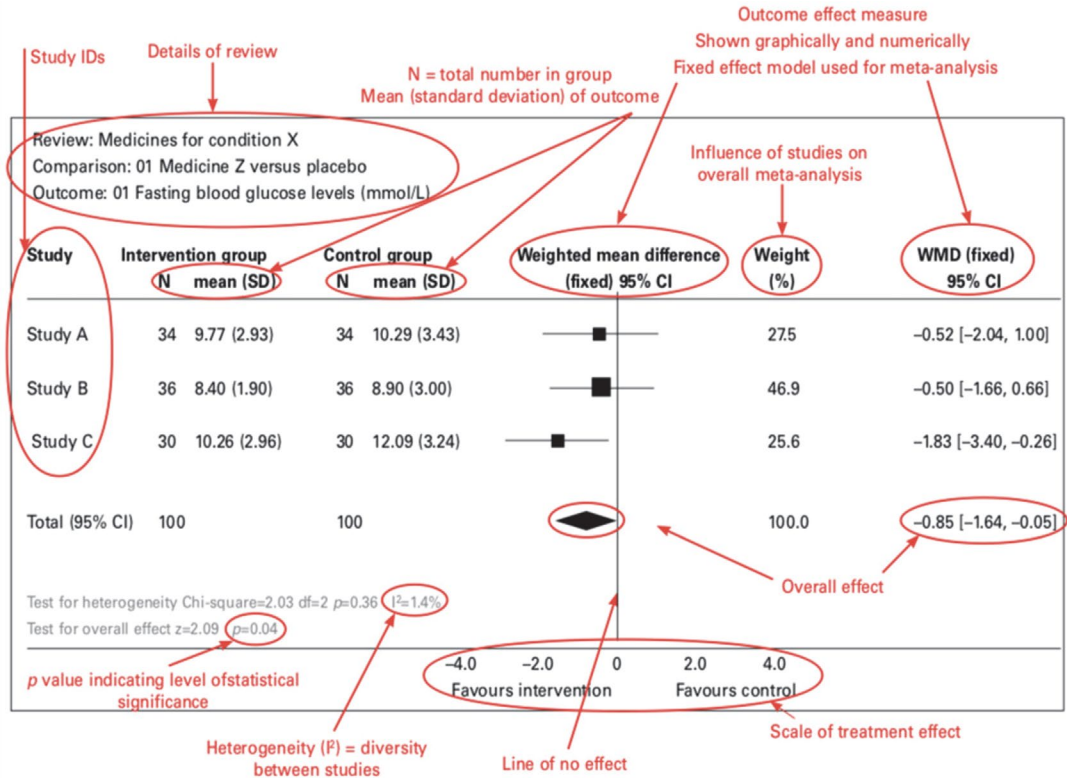


Fig. 16.3 Forest plot elements, labeled (Ried 2006) [35]

effect estimate and its associated whiskers represent the confidence interval (CI). The longer the whiskers, the less precise the estimate. If the CI crosses the “line of no effect,” it is considered not statistically significant. This is also expressed as a probability value (p value) in the “test for overall effect.” The weighting of an individual study is represented by the size of its box and is typically determined by its sample size and the precision of its results [35]. Occasionally, studies are also weighted by their methodological quality [1]. At the bottom of the Forest plot, the data are pooled from the weighted averages of the study results and is depicted as a diamond, where the centre of the diamond indicates the overall effect estimate and the width of the diamond depicts the overall CI [35].

Finally, the pooled effect size should ideally be represented in units that are easy to interpret and meaningful to clinicians and patients. Dichotomous outcomes should be reported as absolute risk reduction and number needed to treat (see Chap. 6). Continuous outcomes should reference a patient-important effect size, such as minimal important difference [36].

In the Kim et al. study [14], 33 articles were included in qualitative analysis and 29 were selected for meta-analysis, excluding four studies

that did not report numerical data on any outcomes of interest. Eleven of these 29 articles measured our outcome of interest (capsular contracture rate). After excluding one article to eliminate heterogeneity, the calculated pooled odds ratio was 0.37 (95% CI: 0.41, 0.79) (Fig. 16.4), indicating that the ratio of the odds of developing capsular contracture following prepectoral implant placement versus subpectoral placement was 0.37. There was no interpretation of the odds ratio with respect to effect size. The forest plots also do not display a line for an odds ratio of 1 in the middle of the x-axis, which reduces the clarity of the figures.

How Precise Were the Results?

Precision of results is determined by the size of the CI around its point estimate. A small sample size and a large standard deviation both widen the CI. A narrow CI indicates a more precise summary estimate [6].

In the Kim et al. article [14], the CI informs us that we can be certain with 95% probability that the ratio of the odds of developing capsular contracture following prepectoral implant placement versus subpectoral placement lies between 0.25 and 0.56 (Fig. 16.4). This relatively narrow CI indicates a precise result. It does not cross the line of no effect and the p value is <0.001, indicating statistical significance.

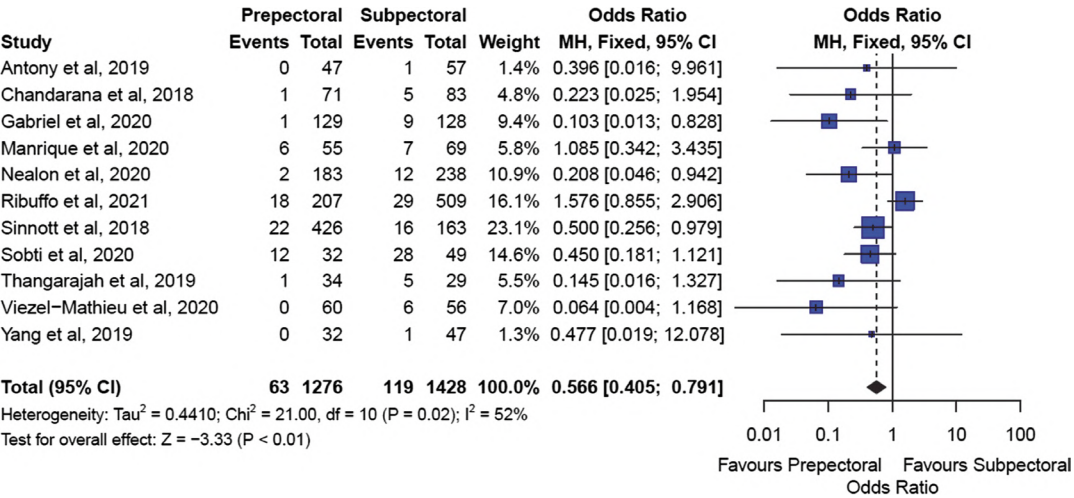


Fig. 16.4 Forest plot for the comparison of capsular contracture. (Adapted from Kim et al. 2016 [14])

Can I Apply the Results to my Patients?

How Can I Best Interpret the Results and Apply Them to My Patients?

While a systematic review or meta-analysis may support the use of a particular intervention, it is the responsibility of the surgeon to ensure that the findings are applicable to his or her patient. After assessing the internal validity of the study, surgeons should review treatment settings, inclusion/exclusion criteria, and participant characteristics to ensure adequate external validity that will demonstrate similar outcomes to their patient population [23].

The study by Kim et al. [14] included English articles published prior to April 2021 that assessed implant-based breast reconstruction with prepectoral and subpectoral techniques. However, the authors do not report any age ranges for inclusion/exclusion. Your patient—a 55-year-old female undergoing implant-based reconstruction, fits the available inclusion criteria. Although Kim et al. [14] did not report the ages of the patients in the included studies, you have no reason to suspect that they would differ drastically from your patient's age.

Were All Patient-Important Outcomes Considered?

Along with a defined primary outcome, other relevant outcomes must be considered when informing clinical decisions. Specifically, the effectiveness of an intervention should be derived by measuring “patient-important outcomes”—the clinical events deemed relevant to a specific patient population [5, 37]. These outcomes may include the occurrence of stroke, death, or quality of life outcomes. For example, potential patient-important outcomes of percutaneous fixation for scaphoid fracture include the risk of infection, reoperation, and neurovascular injury. A focused review of the evidence may provide an accurate result of the effect of fixation on each of these outcomes, but an informed clinical decision requires that all outcomes be considered. As a result, the ideal systematic review will present a series of reviews for each patient-important outcome [6, 22].

A limitation of most systematic reviews and meta-analyses remains that they frequently do not report all relevant patient-important outcomes [1]. While reasons for this may vary, one explanation is that individual studies often measure outcomes differently, or not at all, making it difficult to pool and summarize results [1, 6]. To address this, the implementation of core outcome sets (COS) has been proposed. COS have been defined within the literature as an agreed upon set of outcomes that should be reported at a minimum in all studies assessing a particular condition [38, 39]. This initiative, referred to as Core Outcome Measures in Effectiveness Trials (COMET), seeks to improve homogeneity in outcome reporting to improve outcome-pooling in the form of systematic reviews and meta-analyses [35] (see Chap. 34). Authors and readers of systematic reviews are encouraged to review COMET resources (<https://www.comet-initiative.org/>) to identify whether a set of “core outcomes” exist for their condition of interest.

In addition to capsular contracture rate, Kim et al. [14] also evaluated the rates of infection, hematoma, implant loss, seroma, skin flap necrosis, and wound dehiscence, which are all patient-important outcomes. They found a statistically significant decrease in odds of infection in the prepectoral group compared to the subpectoral group; all other complications did not have statistically significant differences. Postoperative pain, quality of life, and upper arm function were also assessed but there were no quantitative analyses.

Is the Benefit Worth the Potential Costs and Risks?

Either explicitly or implicitly, it is the responsibility of surgeons to make recommendations by weighing the potential costs and benefits of an intervention in the context of the patient's values and preferences [6]. For example, a manual labourer may deem continued observation to be an unacceptable management strategy and may instead opt for prompt surgical intervention. As a result, systematic reviews should avoid specific care recommendations as they are unable to account for specific patient circumstances and

values that can only be determined through direct consultation with the patient [23].

In the chosen example, the patient's primary concern was risk of capsular contracture. Risks of other complications and other patient-important outcomes such as quality of life and satisfaction should be discussed. In non-single payer healthcare systems, the difference in cost between interventions should also be discussed. Overall, given that prepectoral implant placement significantly decreases the odds of capsular contracture without significant differences in other complications, we can infer based on available evidence that prepectoral breast reconstruction may be suitable for this patient. If there was evidence of significant drawbacks to prepectoral implant placement, the evidence should be considered in the context of the patient's own beliefs and values.

Resolution of the Clinical Scenario

At the follow-up appointment, you report your findings to the patient. You conclude that the statistically significant results presented in the study by Kim et al. [14] have adequate internal and external validity and ultimately suggest prepectoral implant placement for her breast reconstruction. You inform the patient that there was also no significant difference in the rate of other complications with prepectoral implant placement. The patient is pleased with this explanation and the decision is made to pursue prepectoral implant-based breast reconstruction.

Conclusion

While systematic reviews and meta-analyses represent the top of the level of evidence hierarchy, they are not without limitations. Consequently, surgeons must possess the skills to interpret the results of review articles to recognize the potential for bias and appropriately assess their application to clinical practice.

Appendix A—Search Results

1. Alnaif N, Safran T, Viesel-Mathieu A, Alhalabi B, Dionisopoulos T. Treatment of breast animation deformity: A systematic review. *J Plast Reconstr Aesthet Surg*. 2019;72 (5):781–8.
2. Awadeen A, Fareed M, Elameen AM. Postoperative complications following prepectoral versus partial subpectoral implant-based breast reconstruction using ADM: a systematic review and meta-analysis. *Aesthet Plastic Surg*. 2023;47(1):81–91.
3. Ching AH, Lim K, Sze PW, Ooi A. Quality of life, pain of prepectoral and subpectoral implant-based breast reconstruction with a discussion on cost: A systematic review and meta-analysis. *J Plast Reconstr Aesthet Surg*. 2022;75(8):2550–60.
4. Fusario D, Casella D. Postoperative complications following prepectoral versus partial subpectoral implant-based breast reconstruction using ADM: a systematic review and meta-analysis. *Aesthet Plastic Surg*. 2024;48(1):13–4.
5. Kim YH, Yang YJ, Lee DW, Song SY, Lew DH, Yang EJ. Prevention of postoperative complications by prepectoral versus subpectoral breast reconstruction: a systematic review and meta-analysis. *Plast Reconstr Surg*. 2024;153(1):10e–24e.
6. Li L, Su Y, Xiu B, Huang X, Chi W, Hou J, Zhang Y, Tian J, Wang J, Wu J. Comparison of prepectoral and subpectoral breast reconstruction after mastectomies: A systematic review and meta analysis. *Eur J Surg Oncol*. 2019;45(9):1542–50.
7. Liu J, Zheng X, Lin S, Han H, Xu C. A systematic review and meta-analysis on the prepectoral single-stage breast reconstruction. *Support Care Cancer*. 2022;30(7):5659–68.
8. Montorfano L, Hung YC, Chaker S, Saad M, Kalmar CL, Ferri F, Higdon KK, Perdakis G. Examination of outcome disparities in reports of prepectoral and subpectoral direct-

to-implant reconstruction: a systematic review and meta-analysis. *Ann Plastic Surg.* 2023;90(5):506–15.

9. Ostapenko E, Nixdorf L, Devyatko Y, Exner R, Wimmer K, Fitzal F. Prepectoral versus subpectoral implant-based breast reconstruction: a systemic review and meta-analysis. *Ann Surg Oncol.* 2023;30(1):126–36.
10. Pagliara D, Schiavone L, Garganese G, Bove S, Montella RA, Costantini M, Rinaldi PM, Bottosso S, Grieco F, Rubino C, Salgarello M, Ribuffo D. Predicting mastectomy skin flap necrosis: a systematic review of preoperative and intraoperative assessment techniques. *Clin Breast Cancer.* 2023;23(3):249–54.
11. Zhu L, Liu C. Postoperative complications following prepectoral versus partial subpectoral implant-based breast reconstruction using ADM: a systematic review and meta-analysis. *Aesthet Plastic Surg.* 2023;47(4):1260–73.
12. Zhu L, Liu C. Invited Response on: Postoperative complications following prepectoral versus partial subpectoral implant-based breast reconstruction using ADM: a systematic review and meta-analysis. *Aesthet Plastic Surg.* 2024;48(1):15–6.

Appendix B—Search Terms [13]

((((Mastectomies [Title/Abstract] OR Mastectomy [Title/Abstract] OR Mastectomy, skin sparing [Title/Abstract] OR Mastectomy, Nipple sparing [Title/Abstract] OR (((“Mastectomy”[Mesh]) OR “Mastectomy, Simple” [Mesh]) OR (“Mastectomy, Segmental” [Mesh] OR “Mastectomy Modified Radical” [Mesh] OR “Mastectomy, Extended Radical” [Mesh] OR “Mastectomy, Subcutaneous” [Mesh] OR “Mastectomy, Radical” [Mesh] OR “Prophylactic

Mastectomy” [Mesh]))) AND ((Breast reconstruction [Title/Abstract] OR Breast reconstructions [Title/Abstract] OR Prepectoral [Title/Abstract] OR Subcutaneous [Title/Abstract] OR Subpectoral [Title/Abstract] OR Implant [Title/Abstract] OR

Prosthesis [Title/Abstract] OR Prosthetic [Title/Abstract] OR (“Tissue Expansion Devices” [Mesh])) AND ((Radiation [Title/Abstract] OR Radiotherapy [Title/Abstract] OR Radiation Therapy [Title/Abstract] OR (“Radiation” [Mesh] OR “Radiotherapy” [Mesh])).

References

1. Bhandari M, Devereaux P. J., Montori V, Cinà C, Tandan V, Guyatt GH, The Evidence-Based Surgery Working Group. Users' guide to the surgical literature: how to use a systematic literature review and meta-analysis. *Can J Surg.* 2004;47(1):60–7.
2. Haines T, McKnight L, Duku EK, Perry L, Thoma A. The role of systematic reviews in clinical research and practice. *Clin Plast Surg.* 2008;35(2):207–14.
3. Kelley BP, Chung KC. Developing, conducting, and publishing appropriate systematic review and meta-analysis articles. *Plast Reconstr Surg.* 2018;141(2):516–25.
4. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics.* 1994;50(4):1088–101.
5. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med.* 1996;125(7):605–13.
6. Guyatt GH, Rennie D, Meade MO, Cook DJ, editors. Users' guides to the medical literature: a manual for evidence-based clinical practice. New York: McGraw-Hill Education; 2015.
7. Samargandi OA, Hasan H, Thoma A. Methodologic quality of systematic reviews published in the plastic and reconstructive surgery literature: a systematic review. *Plast Reconstr Surg.* 2016;137(1):225e–36e.
8. Higgins JP, Thomas J, editors. Cochrane handbook for systematic reviews of interventions (Version 6.4). London: The Cochrane Collaboration; 2023.
9. Sprague S, McKay P, Thoma A. Study design and hierarchy of evidence for surgical decision making. *Clin Plast Surg.* 2008;35(2):195–205.
10. Egger M, Smith GD, Sterne JA. Uses and abuses of meta-analysis. *Clin Med.* 2001;1(6):478–84.
11. Petrisor BA, Bhandari M. The hierarchy of evidence: levels and grades of recommendation. *Indian J Orthop.* 2007;41(1):11–5.
12. Zhu L, Liu C. Postoperative complications following prepectoral versus partial subpectoral implant-based breast reconstruction using ADM: a systematic review and meta-analysis. *Aesth Plast Surg.* 2023;47(4):1260–73.
13. Ostapenko E, Nixdorf L, Devyatko Y, Exner R, Wimmer K, Fitzal F. Prepectoral versus subpectoral implant-based breast reconstruction: a systemic review and meta-analysis. *Ann Surg Oncol.* 2023;30(1):126–36.

14. Kim YH, Yang YJ, Lee DW, Song SY, Lew DH, Yang EJ. Prevention of postoperative complications by Prepectoral versus subpectoral breast reconstruction: a systematic review and meta-analysis. *Plast Reconstr Surg.* 2024;153(1):10e–24e.
15. Thoma A, Eaves FF III. What is wrong with systematic reviews and meta-analyses: if you want the right answer, ask the right question! *Aesthet Surg J.* 2016;36(10):1198–201.
16. Thoma A, McKnight L, McKay P, Haines T. Forming the research question. *Clin Plast Surg.* 2008;35(2):189–93.
17. Waltho D, Kaur MN, Haynes RB, Farrokhyar F, Thoma A. Users' guide to the surgical literature: how to perform a high-quality literature search. *Can J Surg.* 2015;58(5):349–58.
18. Zlowodzki M, Poolman RW, Kerkhoffs GM, Tornetta P III, Bhandari M, International Evidence-Based Orthopedic Surgery Working Group. How to interpret a meta-analysis and judge its value as a guide for clinical practice. *Acta orthopaedica.* 2007;78(5):598–609.
19. Lane D, Dykeman J, Ferri M, Goldsmith CH, Stelfox H. Capture-mark-recapture as a tool for estimating the number of articles available for systematic reviews in critical care medicine. *J Crit Care.* 2013;28(4):469–75.
20. Montori VM, Smieja M, Guyatt GH. Publication bias: a brief review for clinicians. *Mayo Clin Proc.* 2007;75(12):1284–8.
21. Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan AR, McIlroy W, Oxman AD. Should unpublished data be included in meta-analyses?: Current convictions and controversies. *JAMA.* 1993;269(21):2749–53.
22. Montori VM, Swiontkowski MF, Cook DJ. Methodologic issues in systematic reviews and meta-analyses. *Clin Orthop Relat Res.* 2003;413:43–54. <https://doi.org/10.1097/01.blo.0000079322.41006.5b>.
23. Tseng TY, Dahm P, Poolman RW, Preminger GM, Canales BJ, Montori VM. How to use a systematic literature review and meta-analysis. *J Urol.* 2008;180(4):1249–56.
24. Booth A, Clarke M, Dooley G, Ghera D, Moher D, Petticrew M, Stewart L. PROSPERO at one year: an evaluation of its utility. *Syst Rev.* 2013;2(1):1–7.
25. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 2009;6(7):e1000097.
26. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, Moher D, Tugwell P, Welch V, Kristjansson E. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ.* 2017;358:j4008.
27. Jadad A, Moore R, Carroll D, Jenkinson C, Reynolds D, Gavaghan D, McQuay HJ. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials.* 1996;17(1):1–12.
28. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. *ANZ J Surg.* 2003;73(9):712–6.
29. Voineskos SH, Coroneos CJ, Ziolkowski NI, Kaur MN, Banfield L, Meade MO, Thoma A, Chung KC, Bhandari M. A systematic review of surgical randomized controlled trials: part I. Risk of bias and outcomes common pitfalls plastic surgeons can overcome. *Plast Reconstr Surg.* 2016;137(2):696–706.
30. Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. *BMJ.* 2001;323(7303):42–6.
31. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA.* 1999;282(11):1054–60.
32. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials.* 1995;16(1):62–73.
33. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med.* 2009;6(7):e1000100.
34. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327(7414):557.
35. Ried K. Interpreting and understanding meta-analysis graphs: a practical guide. *Aust Fam Physician.* 2006;35(8):635–8.
36. Evaniew N, van der Watt L, Bhandari M, Ghert M, Aleem I, Drew B, Guyatt GH. Strategies to improve the credibility of meta-analyses in spine surgery: a systematic survey. *Spine J.* 2015;15(9):2066–76.
37. Gallo L, Eskicioglu C, Braga LH, Farrokhyar F, Thoma A. Users' guide to the surgical literature: how to assess an article using surrogate end points. *Can J Surg.* 2017;60(4):280.
38. Prinsen CA, Vohra S, Rose MR, King-Jones S, Ishaque S, Bhaloo Z, Adams D, Terwee CB. Core Outcome Measures in Effectiveness Trials (COMET) initiative: protocol for an international Delphi study to achieve consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set'. *Trials.* 2014;15:247.
39. Potter S, Holcombe C, Ward JA, Blazeby JM, BRAVO Steering Group. Development of a core outcome set for research and audit studies in reconstructive breast surgery. *Br J Surg.* 2015;102(11):1360–71.

Prospective and Retrospective Cohort Studies

17

Richard Walker, Ramy Behman, Lev Bubis,
and Paul Karanicolas

Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
CI	Confidence Interval
DAG	Directed Acyclic Graph
HIV	Human Immunodeficiency Virus
RCT	Randomized Controlled Trial

Introduction

Cohort studies are a type of observational study that follows samples of at-risk individuals forward in time to examine associations between measured exposures and future outcomes. Exposures may include surgical techniques,

medical interventions, patient characteristics, or health-system factors [1].

Cohort studies can be categorized as prospective or retrospective. In prospective cohort studies, enrolment of the cohort and determination of exposure status occur before any participants experience the outcome with participants subsequently followed forward in real time. Retrospective cohort studies, by contrast, are designed after outcomes have occurred. Investigators conducting retrospective cohort studies look backward in time to identify a cohort of at-risk subjects who are representative of a target population, ascertain the exposure status, then determine if they experienced the outcome. In general, retrospective cohort studies are less costly and time consuming compared to prospective studies. However, retrospective cohort studies often rely on data sources designed for other purposes. These sources may not be as complete or accurate compared to those created for prospective studies.

Compared to cohort studies, randomized controlled trials (RCTs) generally provide a higher level of evidence about the effectiveness of interventions owing to their lower risk of bias [2]. Case-control studies are generally better suited than cohort studies for investigating rare outcomes since cases in case-control studies are selected based on the occurrence of outcomes. Nevertheless, cohort studies have an essential role in evidence-based surgery. Cohort studies

R. Walker
Department of Surgery, University of Toronto,
Toronto, ON, Canada
e-mail: richard.walker@mail.utoronto.ca

R. Behman
Division of Surgery, University of Texas MD
Anderson Cancer Center, Houston, TX, USA
e-mail: rsbehman@mdanderson.org

L. Bubis
Department of Surgery, University of Manitoba,
Health Sciences Centre, Winnipeg, MB, Canada
e-mail: lbubis@hsc.mb.ca

P. Karanicolas (✉)
Department of Surgery, University of Toronto,
Sunnybrook Health Sciences Centre, Sunnybrook
Research Institute, Toronto, ON, Canada
e-mail: paul.karanicolas@sunnybrook.ca

are the optimal study design to evaluate prognostic factors; they allow for the efficient longitudinal follow-up of large, representative study samples, including subgroups often excluded or under-sampled in RCTs [3–5]. Cohort studies are also invaluable for the study of multiple exposures, rare exposures, exposures with known risks to which it would be unethical to allocate participants in a trial, and in situations where there is a long interval of time between exposure and outcome.

Clinical Scenario

You are the resident in a busy general surgery clinic, where you are seeing a 67-year-old woman in consultation for biliary colic. She describes no previous episodes of cholecystitis and has a surgical history significant only for an open hysterectomy for uterine fibroids and dyslipidemia. She is interested in an elective cholecystectomy to address her recurrent postprandial abdominal pain. After discussing the risks and benefits of laparoscopic cholecystectomy, she tells you she has been reading in the news about robot-assisted surgery, and asks whether a robot-assisted approach would offer any advantages over conventional laparoscopic surgery. You tell her you are unsure if there are advantages. She tells you that a relative needed a reoperation for an injury to the bile ducts, so she would like to think about surgery and would like to see you again in another few weeks to revisit the discussion. You decide you are going to come prepared for her next appointment with the latest evidence for robot-assisted cholecystectomy.

Literature Search

That evening (March 25, 2024), you search Medline for “Robotic Surgical Procedures AND Cholecystectomy, Laparoscopic.” You find a recently published manuscript titled “Comparative Safety of Robotic-Assisted vs Laparoscopic Cholecystectomy” [6]. You decide to begin your literature review with this article, as you believe its large sample size, recruitment across multiple

institutions, and recency of completion make it likely to be relevant to your patient.

Article Summary

The article by Kalata and colleagues describes a multicenter retrospective cohort study comparing robot-assisted and laparoscopic cholecystectomy regarding bile duct injury rates and 30-day complications [6]. In a cohort of 1,026,088 Medicare beneficiaries between 66 to 99 years of age, robot-assisted cholecystectomy was associated with an increased risk of bile duct injury necessitating operative repair. Robot-assisted surgery was also associated with a higher rate of postoperative biliary interventions such as endoscopic stenting. There was no difference in overall 30-day complications between the two approaches. An outline of the study by Kalata and colleagues is provided in Table 17.1.

Using the study by Kalata and colleagues [6] as a case study, this chapter will present a practical guide for the appraisal of cohort studies in surgery (Box 17.1). Our aim is to provide practicing surgeons and trainees with a framework for the evaluation of the methodological quality and applicability of cohort studies.

Box 17.1 Guideto Interpretation of Cohort Studies in Surgery [7]

1. *What are the study’s research question and clinical relevance?*
2. *Are the results valid?*
 - I. What is the study population?
 - II. Are the data sources reliable?
 - III. Are the data collection methods sound?
 - IV. Do the authors provide clear and justifiable definitions of the primary and secondary outcomes?
 - V. Have the authors accounted for potential sources of bias?
 - VI. Is the statistical analysis (a) valid and (b) described clearly?

3. *What are the results?*

- I. Are baseline characteristics of the exposure groups clearly reported and accounted for?
- II. Are the results and relevant statistics clearly stated?
- III. How precise were estimates of effect size?
- IV. Are results robust to planned sensitivity analyses?
- V. Were subgroup analyses used? If so, were these specified a priori and is the issue of multiple testing addressed?

4. *How can I interpret these results in the context of my clinical practice?*

- I. Are the results of the study placed in the broader context of the clinical question?
- II. Is there a description of the existing literature?
- III. Were all clinically important outcomes considered?
- IV. Are limitations of the study clearly described?

This list was modified from Thoma et al. [7].

What Are the Study's Research Question and Clinical Relevance?

Study objectives and hypotheses should be described clearly and supported by a coherent rationale, which outlines the clinical importance of the research question. It is important to note that cohort studies are vulnerable to the problematic practices of data dredging and selective reporting [8, 9]. When reading a cohort study, it is often impossible to know whether research questions or methods were altered during the research conduct or if additional unreported analyses were undertaken. As such, protocol registration prior to the commencement of research bolsters the validity of studies. While journals do not currently

require authors to register protocols for observational studies, registration is encouraged [10].

In the article under review, Kalata and colleagues [6] appropriately described the rationale for conducting their study, including the importance of the research question and the limits of existing evidence. The authors describe the expanding use of robot-assisted surgery in the United States despite a lack of evidence demonstrating safety for some procedures.

Are the Results Valid?

What Is the Study Population?

The population of interest should be explicitly specified. A detailed description of the inclusion and exclusion criteria for cohort entry is essential for readers to evaluate to whom the study results apply. Sample size calculations based on anticipated effect sizes in comparative effectiveness studies are valuable to ensure studies are adequately powered to detect clinically meaningful difference between interventions, should they truly exist. However, poor statistical power due to a low sample size should not preclude the conduct of observational studies using existing data sources, such as retrospective cohort studies. This point is eloquently discussed by Miguel Hernán in his 2022 article for the *Journal of Clinical Epidemiology* [11]. Hernán argues that while poor statistical power results in effect estimates with low precision, these estimates can still be used in meta-analyses to calculate pooled estimates with improved precision. Having several imprecise estimates is better than having no estimates at all. As well, there is little potential harm to patients when using existing, anonymized databases for retrospective studies. Therefore, while high statistical power is ideal, it is not necessary for retrospective cohort studies. On the other hand, an adequate sample size is very important for experimental studies such as RCTs. In these studies, participants are exposed to risks by virtue of participation. It would be unethical to enrol participants and expose them to risks if the study is inadequately powered to answer the research question.

Table 17.1 Outline of the retrospective cohort study conducted by Kalata et al. [6]

Study element	Summary
Population and setting	1,026,088 patients aged 66 to 99 years with Medicare fee-for-service coverage who underwent a robot-assisted or laparoscopic cholecystectomy at a nonfederal acute care hospital between January 1, 2010 and December 31, 2019. Patients were excluded if they underwent surgery for cancer.
Exposure/comparators and allocation	<i>Exposure</i> —robot-assisted cholecystectomy, defined using International Classification of Diseases (ICD) codes. <i>Comparator</i> —conventional laparoscopic cholecystectomy <i>Allocation</i> —selection of the operative approach was at the discretion of the treating surgeon
Results	<i>Primary outcome:</i> <i>Adjusted rate of bile duct injury requiring surgical intervention with hepaticojejunostomy or choledocojejunostomy within 1 year</i> —0.7% in the robot-assisted group versus 0.2% in the laparoscopic group, corresponding to a relative risk of 3.16 (95% CI 2.57 to 3.75) <i>Secondary outcomes:</i> <i>Adjusted rate of biliary intervention (including endoscopic retrograde cholangiopancreatography and common bile duct exploration)</i> —7.4% in the robot-assisted group versus 6.0% in the laparoscopic group, corresponding to a relative risk of 1.25 (95% CI 1.16 to 1.33) <i>Adjusted rate of 30-day overall complications</i> —20.5% in the robot-assisted group versus 20.6% in the laparoscopic group, corresponding to a relative risk of 1.00 (95% CI 0.97 to 1.03) <i>Adjusted rate of serious complications requiring a length of stay > 75th percentile</i> —9.3% in the robot-assisted group versus 8.6% in the laparoscopic group, corresponding to a relative risk of 1.08 (95% CI 1.03 to 1.13)
Authors' conclusions	Robot-assisted cholecystectomy was associated with higher rates of bile duct injury compared to the conventional laparoscopic approach. In the absence of other advantages, the data call into question the use of the robotic platform for cholecystectomy

Abbreviations: *CI* confidence interval

In their article, Kalata et al. [6] appropriately describe study inclusion and exclusion criteria. The inclusion of Medicare beneficiaries across the United States makes the results generalizable to a broad population. Of note, the study does not include patients aged 65 and younger. The authors make the compelling argument that cases are managed similarly regardless of age, so they would not expect the results to differ for a younger population. A sample size calculation is not included in the manuscript; however, the inclusion of over 1 million patients makes the study well suited for evaluating uncommon outcomes.

Are the Data Sources Reliable?

There should be a clear description of data sources. In retrospective cohort studies, data may be obtained from primary sources (i.e., from chart abstraction and prospective institutional databases) or secondary sources designed for alternate purposes (i.e., from administrative, demographic, and physician claims databases)

[12]. Some studies have demonstrated poor accuracy of secondary data for surgical outcomes assessment [13, 14]. However, retrospective chart abstraction or the creation of institutional databases require a large amount of resources and are unfeasible for large, population-based studies. For research using administrative databases, validity may be bolstered through the conduct of preliminary studies assessing the accuracy of the databases [15, 16].

In prospective studies, data may be obtained via active data collection processes specific to the study or through linkage to routinely collected data from external sources such as registries or population-based databases.

Kalata et al. [6] abstracted data from the Medicare Provider Analysis and Review File for their study. This administrative database contains hospital stay records for Medicare beneficiaries. Administrative datasets are often limited by measurement error. The authors attempted to mitigate this issue by selecting diagnoses and outcomes

that are reliably coded in these databases. The authors note the overall rates of bile duct injury in their study are similar to those from other cohorts, suggesting they were able to capture their primary outcome with reasonable accuracy.

Another limitation of administrative databases is the unavailability of potentially important covariates. The authors acknowledge they were unable to measure the severity and chronicity of gallbladder inflammation, which is a risk factor for bile duct injury. Unmeasured covariates have the potential to confound associations in cohort studies.

Are the Data Collection Methods Sound?

Authors should report on the methods of data collection as well as the extent of missing data. The reliability of cohort studies may be bolstered by creating clear data collection plans inclusive of detailed variable definitions prior to study conduct. Data collection plans can be included in study protocols and reassure readers that results were not being selectively reported and variable definitions were not being adjusted to find favourable results.

Missing data may compromise the validity of study results. In a systematic review of RCTs published in high impact journals over 2 years, significant results became non-significant in up to 33% of trials after making plausible assumptions about event rates in participants lost to follow-up [17]. As such, considerable effort should be made to minimize missing data and the method of handling missing data in statistical analyses should be described. In addition, where multiple contributors participate in data collection, assessment of consistency (e.g., inter-rater reliability) is important to assure the quality of data collection [18].

Kalata et al. [6] provide coding definitions for several covariates and outcomes used in their study. Coding definitions for 30-day complications were not provided. Minor complications such as superficial wound infections can be difficult to capture in administrative databases. The authors use hospital length of stay >75th percentile as a surrogate for serious complications. While this is a convenient measure of severity, major complications such as death may still

result in hospital stays less than the 75th percentile in duration. Missing data is not reported.

Do the Authors Provide Clear and Justifiable Definitions of the Primary and Secondary Outcomes?

The outcomes measured in surgical cohort studies should address the key aspects of the clinical question under study. When applicable, authors should justify their designation of primary and secondary outcomes. Methods of outcome ascertainment must be carefully planned and reported with sufficient clarity to be replicable. Kalata et al. [6] explicitly specify their primary and secondary outcomes. The primary outcome was bile duct injury necessitating surgical repair with hepaticojejunostomy or choledocojejunostomy within 1 year after cholecystectomy. The authors justify the selection of this primary outcome as clinically important and reliably measured in their dataset. Of note, this primary outcome excludes bile duct injuries that do not require hepaticojejunostomy or choledocojejunostomy to repair. Therefore, some clinically important information may be missed in the assessment of the primary outcome. The authors do, however, include bile duct injuries requiring endoscopic management or operative bile duct exploration as a secondary outcome.

In studies of surgical interventions, the time frame of outcome assessments is a key consideration. A high proportion of postoperative complications are diagnosed only after hospital discharge. Data sources that are limited to the index hospitalization may, therefore, substantially undercount postoperative adverse events [13, 19]. This may lead to biased results when study groups have differing risks of short-term (i.e., in-hospital) and longer-term (i.e., 30-day) complications. Therefore, a clear description of timeframes for outcome assessment is imperative to allow readers to fully understand the implications of the results. Kalata et al. clearly describe the time frame for outcome assessment [6].

The number of outcomes being assessed may also be worth noting. Studies should have well-justified outcomes of interest, and ideally not

assess too many outcomes. As the number of outcomes being assessed increases, so does the probability of encountering at least one false-positive result. The concept of evaluating many outcomes and increasing the risk of a false positive result is termed the multiple comparisons problem.

For example, consider a hypothetical study comparing apples from two orchards. A batch of apples is randomly selected from each orchard's harvest. Even if the orchards are identical, by random chance, there may be some differences between the batches. The probability of finding a significant difference between the batches by chance alone increases as the number of comparisons increases. It is therefore important for studies to have a justification for evaluating each outcome, and for readers to be wary of significant results if many outcomes are assessed.

In studies with many statistical tests, there are methods that can be employed to account for the increased risk of a false-positive result. Essentially, these methods raise the bar for how large a difference must be for it to be considered significant. One widely known and straightforward method is the Bonferroni correction. Instead of using a significance level of $p \leq 0.05$, the significance level is divided by the number of tests being performed [20, 21]. So, ≤ 0.005 to be considered statistically significant. Of note, the Bonferroni correction is considered to be highly conservative and has been criticized for unnecessarily increasing the risk of false negative results [22]. Adjustments for multiple comparisons should therefore be used with careful consideration of the relative importance of avoiding a false-positive versus a false-negative result, whether multiple tests are being performed without justification, and whether the universal null hypothesis that all tests are not significant is important to evaluate [21].

Have the Authors Accounted for Potential Sources of Bias?

Bias in observational studies refers to the occurrence of systematic errors in study design or conduct. Compared to random error, which is variable and unpredictable, bias results in sys-

tematic deviations from the estimation of true values. The internal validity of a study denotes the extent to which its causal conclusions are justified and rests upon the minimization of bias. There are myriad subtypes of bias [23, 24]. In general, sources of bias can be classified into the following categories: (a) selection bias, (b) information bias, and (c) confounding [25, 26].

Selection Bias

Selection bias refers to systematic errors that arise from the selection of patients for inclusion in the study. Selection bias may occur when factors related to inclusion of study subjects are associated with exposures and outcomes leading to distortion of effect estimates or unrepresentativeness of the study sample relative to the target population [27]. One simple example is a theoretical study where sampled patients must either be taking a statin or have undergone a knee arthroplasty. Because of these selection criteria, an analysis would identify an inverse association between statin use and the risk of osteoarthritis in the study population (participants taking a statin are less likely to have undergone a knee arthroplasty for osteoarthritis than those who are not taking a statin). While this example can be hard to imagine occurring in practice, selection of participants may have less obvious associations with the exposure and outcome of interest, and distorted results may occur insidiously. Investigators need to be careful not to create these false associations by virtue of how patients are selected for inclusion.

Selection bias can occur due to differential loss to follow-up in cohort studies. Hernán et al. [28] present an example of a study evaluating the effect of an antiretroviral (the exposure) on the risk of AIDS (the outcome) in persons living with HIV. In this example, side effects from antiretrovirals may cause a study participant to miss follow-up visits. Symptoms from AIDS can also result in loss to follow-up. Patients in the antiretroviral group with complete follow-up are less likely to have a second risk factor for loss to follow-up (i.e., AIDS). Restricting the analysis to

those with complete follow-up creates an association between antiretroviral therapy and AIDS risk, even if there was not a causal relationship.

Cohort studies in surgery may be especially prone to certain types of selection bias. Immortal person-time bias occurs in instances where exposure status is assigned after the onset of follow-up [29]. Assignment to one exposure group, therefore, may be conditional on survival from the time of study onset [30]. Participants who survived to the time of exposure assignment may no longer be representative of the sampled population, leading to biased associations. Confounding by indication is another important source of bias in cohort studies comparing alternative treatments. Confounding by indication refers to situations where treatment group allocation is dependent upon prognostic factors, which may not all be measured in a cohort study [31].

In the study reported by Kalata et al. [6], allocation to study groups was likely based on surgeon selection, institution, and perhaps also patient preference in the case of elective cholecystectomy. While measured baseline variables of the two study groups were adjusted for in a multivariable analysis, unmeasured patient- and disease-related factors may influence allocation by surgeon judgment, leading to selection bias. In contrast to RCTs, in which random allocation produces balance between measured and *unmeasured* covariates, the problem of selection bias is never wholly surmountable in cohort studies. Readers must judge whether the most important covariates for a clinical question are measured and addressed.

Information Bias

Information bias refers to systematic errors in the classification or measurement of variables. Sources of information bias may be the observer, subject, or measurement instrument. Information bias is differential if the error in variable classification is dependent on the value of the outcome or exposure of interest; otherwise, it is nondifferential [26]. Information bias should be nondifferential in cohort studies if consistent data

collection practices are applied across all study participant. However, in practice, this may not be the case. Two examples of differential information bias are observer bias and interviewer bias.

Observer bias occurs when outcome ascertainment is influenced by assessors' knowledge of subjects' exposures. Observer bias can be minimized by blinding outcome assessors, having multiple outcome assessors review each participant, using objective outcome measures and by allowing differential levels of certainty for outcome ascertainment (i.e., outcome "possible," "probable," or "definite") [27]. Interviewer bias occurs in prospective studies using personalized interviewers. If interviewers are unblinded to exposure status, they risk consciously or subconsciously obtaining answers which accord with their preconceptions [27]. In addition to ensuring blinding of study interviewers, measures such as standardized data collection forms and interview scripts may reduce interviewer bias.

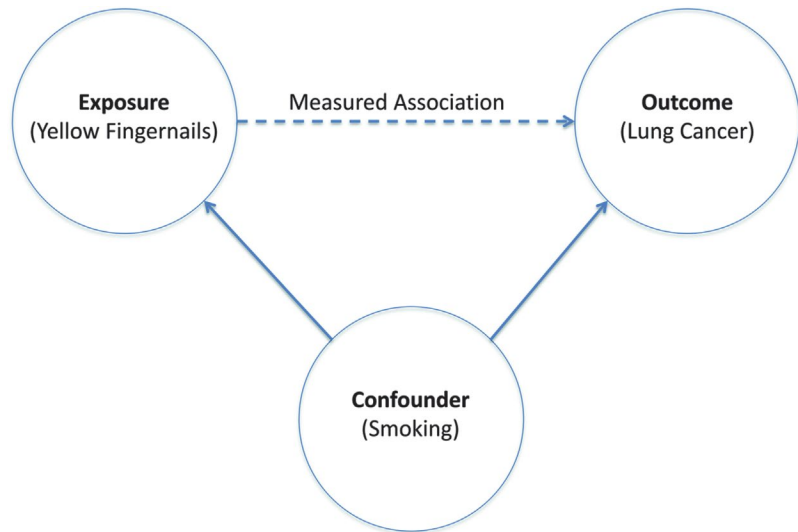
Kalata et al. [6] used an objective primary outcome (bile duct injury requiring hepaticojunostomy or choledocojunostomy). However, divergent methods of outcome assessment at alternate sites in this multicenter study may lead to bias through differential ascertainment of outcomes. This is especially relevant for outcomes that are more challenging to measure, such as the components of 30-day complications (e.g., pneumonia, surgical site infection, and delirium).

Confounding

Confounding occurs when an apparent association between two factors is distorted by extraneous factors [32]. This distortion occurs when there is an extraneous factor that is independently associated with both the exposure and the outcome of interest [32, 33]. The effect of an exposure on an outcome may be mixed with or mistaken for the effect of an extraneous variable that has not been accounted for (Fig. 17.1).

A classic example of confounding is the association between yellow fingernails and lung cancer. A confounder (smoking) is associated with both the exposure (yellow fingernails) and the

Fig. 17.1 Directed acyclic graph (DAG) illustrating confounding of observed relationship between yellow fingernails and lung cancer by an extraneous factor (smoking)



outcome (lung cancer). Without appropriately identifying and accounting for potential confounding, a distorted relationship between the exposure and the outcome (that yellow fingernails cause lung cancer) may have been concluded. Similarly, studies of surgical outcomes must consider potential confounders of associations between surgical interventions and clinical outcomes.

Confounding may be divided into two groups: measured confounding and unmeasured confounding [34]. Measured confounding describes potential factors that were measured in the study that can be accounted for in the analysis. Unmeasured confounding refers to potential confounders that were not captured in data collection.

In cohort studies, accounting for potential confounding is often limited to the analysis phase through a variety of statistical approaches. Broadly, these approaches include matching (including propensity score techniques), stratification by potential confounders, and regression (e.g., multivariable models) [35–37]. In the article by Kalata et al. [6], the authors use multivariable regression to estimate the association between surgical approach and outcomes of interest while adjusting for potential confounders. These confounders included age, race and ethnicity, sex, comorbidities, primary diagnosis, and year. This study did not adjust for other potentially important confounders such as surgeon

experience and case complexity. It is possible that inexperienced surgeons were using the robotic platform for cholecystectomy to practice using the technology.

Of note, a confounder should not be confused with a mediator. A variable is a mediator when it contributes to the causal pathway between exposure and outcome. Consider the causal relationship between motor vehicle collisions on the highway and severity of injury when compared to collisions on city streets. This causal pathway is likely mediated by the speed of collision. If a study comparing highway collisions and city collisions controlled for speed (a mediator), the causal relationship would have been inappropriately obscured. Thus, it is important for clinicians to carefully consider the variables they need to control in observational studies so that observed associations are as close to the truth as possible.

Is the Statistical Analysis (a) Valid and (b) Described Clearly?

A description of the statistical analysis should be provided in sufficient detail that, given the dataset, one could replicate the analysis and find identical results. A clear and detailed description of the statistical analysis not only allows for reproducibility but is also necessary to evaluate the validity of a study's methods.

Readers should consider whether the statistical approaches are (1) appropriate for the variables of interest, and (2) account for potential confounding in a suitable manner. Many considerations are necessary to evaluate the validity of an analytical approach; these are discussed in greater detail in Chaps. 26, 27, 28, 29, and 30.

In the study by Kalata et al. [6], multivariable logistic regression models were used to evaluate associations between robotic surgery and outcomes of interest. The covariates in the models are described and defined. As mentioned previously, the models do not account for surgeon factors such as experience or case complexity. Admittedly, it would be challenging to define the severity and chronicity of gallbladder inflammation, even for the most thorough prospective datasets.

The authors appropriately adjusted for clustering of outcomes within hospital referral regions. Two patients within the same hospital region are more likely to have similar outcomes compared to patients in different hospital regions. Failing to account for this clustering of outcomes within hospital groups can result in a misestimation of standard errors [38].

What Are the Results?

Are Baseline Characteristics of the Exposure Groups Clearly Reported and Accounted For?

A clear description of differences in baseline characteristics between treatment groups provides important contextual information with which subsequent results may be interpreted. In the article under study, the baseline characteristics were different in the two treatment groups [6]. Patients undergoing robot-assisted cholecystectomy were more likely to be younger, female, African American, have a greater comorbidity burden, undergo an elective procedure rather than urgent or emergent, and were less likely to have a diagnosis of cholecystitis. Most of these covariates were appropriately adjusted for in the multivariable analyses. However, readers of observational studies must bear in mind that even

after adjusting for measured baseline characteristics, unmeasured or unreported baseline characteristics may be different between exposure groups.

When baseline characteristics differ clinically between exposure groups, readers should be concerned about selection bias. In these instances, analytic methods such as matching or multivariable regression may be used to account for these differences in baseline characteristics.

Are the Results and Relevant Statistics Clearly Stated?

Results should be clearly described according to the analytical approach outlined in the Methods section of the study. Readers of the study should be able to identify which statistical test yielded the results being described. In addition to the results, the study should report the appropriate statistics (e.g., p value, standardized difference, confidence interval) in a manner that is appropriate to the statistical test and is consistent throughout the study.

In the study by Kalata et al. [6], univariable analyses (comparisons of proportions or means/medians) were reported using p values. Multivariable analyses were reported using absolute differences and relative risks to quantify magnitudes of difference and 95% confidence intervals to quantify precision of estimates. The results are clearly stated, and the reporting is consistent. Readers are easily able to identify a description of the corresponding analytical approach in the Methods section.

How Precise Were Estimates of Effect Size?

Effect estimates may be reported in a variety of ways, depending on the outcomes of interest. Reporting can vary from simple differences in proportions to absolute/relative risk ratios to odds ratios and hazard ratios. Confidence intervals should be reported to allow readers to determine the precision of effect estimates (see Chap. 28).

Confidence intervals provide important information with which effect sizes should be interpreted. Wide confidence intervals suggest a high degree of uncertainty and that effects should be

interpreted with caution; narrow confidence intervals suggest a high degree of certainty of the reported estimate [39, 40]. *P* values, while not specifically reflecting the precision of effect sizes, are a measure of the type I error—the probability that an observed effect or larger is interpreted as statistically different when in truth it is not. Both confidence intervals and *p* values are dependent on sample size and error [39–41].

In the article under study, the authors report both adjusted absolute differences and relative risks for each outcome. Absolute differences provide information about the magnitude of difference and in turn the clinical importance of results. However, absolute differences are limited in that they depend on the baseline risk in the population under investigation. For any given treatment, the absolute difference in a low-risk population will be different than the absolute difference in a high-risk population. Relative risks overcome this limitation by remaining stable across baseline risk groups. Therefore, both absolute differences and relative risks are helpful to report. Kalata et al. [6] report 95% confidence intervals for both absolute differences and relative risks, which provide information regarding the precision of estimates.

Are Results Robust to Planned Sensitivity Analyses?

Sensitivity analyses are used to evaluate the validity of study results by testing whether an effect is consistent under alternate study conditions [42]. By introducing or eliminating some of the inherent uncertainty that existed in the original study design, investigators can test how much of the measured effect size was a result of the independent variables of interest, and how much was a product of uncertainty or potential confounders [42]. For example, when the primary analysis involves evaluating an association using a multivariable model, a sensitivity analysis might evaluate the association using a propensity score method or instrumental variable analysis. If the results of the sensitivity analysis are similar to those of the primary analysis, the credibility of the study conclusions is improved. These methodological decisions should ideally be planned

before the conduct of the study to avoid authors selecting for statistically significant sensitivity analyses (i.e., publication bias).

Kalata et al. [6] reported multiple sensitivity analyses designed to address some of the limitations of their primary analysis. An instrumental variable analysis was performed in an effort to address confounding by unmeasured covariates. Instrumental variables must meet 3 conditions: (i) the variable must have a causal effect on the exposure (or share a common cause with the exposure), (ii) the variable must affect the outcome only by its influence on the exposure, and (iii) the variable must not share common causes with the outcome [43]. If a variable meets all three conditions, the average effect of the exposure on an outcome can be estimated without needing to adjust for confounders. The instrumental variable used by Kalata et al. was the proportional use of robot-assisted cholecystectomy across the hospital referral region in the year prior to a patient's operation [6]. This variable satisfies condition (i) since the use of the robotic platform in a prior year can predict its use in a future case. One concern regarding condition (ii) is that the use of a robotic platform in a prior year may lead to additional experience for the operating surgeon, and therefore influence bile duct injury rates. This may call into question the use of this instrumental variable. However, the authors cite previous studies that used similar variables for evaluating new operative approaches. Condition (iii) is satisfied since the use of the robotic platform in a prior year does not appear to share a common cause with bile duct injuries. Once an instrumental variable is selected, the association between the exposure and outcome can be estimated by dividing the effect of the instrument on the outcome by the effect of the instrument on the exposure [43]. In their instrumental variable analysis, Kalata et al. [6] found robot-assisted cholecystectomy remained associated with a greater risk of bile duct injury. Given the potential violation of condition (ii), there may still be bias from not accounting for surgeon experience.

Kalata et al. [6] acknowledge that a hospital's experience with robot-assisted surgery may con-

found results. They performed a sensitivity analysis stratifying hospitals into robotic volume quartiles. Even in the highest volume quartile, robot-assisted surgery remained associated with a greater risk of bile duct injury. The authors also stratified by clinical urgency, diagnosis, and obesity, acknowledging that cholecystectomies can substantially vary in difficulty. The results of these sensitivity analyses were consistent, strengthening the study's findings

Were Subgroup Analyses Used? If So, Were These Specified *a Priori* and Is the Issue of Multiple Testing Addressed?

Unlike sensitivity analyses, which are performed to test assumptions that have been made by investigators, subgroup analyses are used to determine if the relationship between an exposure and an outcome differs across levels of a third variable [37]. To perform a subgroup analysis, the study population is divided into levels of the third variable and results are compared across groups. The subgroups should be specified prior to the analysis of data, and not be the result of conclusions that are found during data analysis [44]. In the article under study, some of the sensitivity analyses also serve as subgroup analyses [6].

Readers of studies that include subgroup analyses should be aware of issues with multiplicity in which, when multiple subgroup analyses are performed, the probability of false-positive results increase [37]. For more information about subgroup analysis, please see Chap. 30.

How Can I Interpret These Results in the Context of My Clinical Practice?

Are the Results of the Study Placed in the Broader Context of the Clinical Question?

The aim of cohort studies is to use a sample of patients for whom data are available to answer a clinical question that applies to the broader referent patient population. The extent to which results from a study sample may be generalized to the overall population is referred to as the external

validity or applicability of the study. A careful evaluation of study design, setting, and analytical approach is necessary to identify sources of bias that may limit a study's internal validity—a prerequisite for external validity [45].

Investigators should frame their study in the context of a larger clinical question and should clearly indicate how their results may be generalized to the overall population that their cohort aims to represent. The readers of a study then have a responsibility to question whether it is appropriate to do so. Differences in patient population or setting may limit the applicability a study's results to the reader's own practice.

The study by Kalata et al. [6] includes only Medicare beneficiaries aged 66 to 99 years of age. Readers should ask whether the results are generalizable to all cholecystectomy patients, regardless of age. The authors acknowledge this as a potential limitation; however, they discuss that the management and surgical approach is similar regardless of age. Given learning curves and volume outcome associations for surgical procedures, readers should also ask whether the surgeons in the study have similar skillsets to those of the target population. In the case of the study by Kalata et al. [6], procedures were performed in nonfederal acute care hospitals across the United States. This would include a diversity of surgeons, making the results more generalizable to the greater population. However, results may be less generalizable to surgeons with significant training and experience with robot-assisted surgery. These surgeons may have bile duct injury rates that are lower than those observed in the present study.

Is There a Description of the Existing Literature?

A description of the existing literature about a clinical topic and a discussion of how the results of a study fit into the body of literature provides important information to a reader. A study in which the findings are supported by previously published studies increases confidence in the conclusions, as the previously published studies may be performed in different settings and with different patients. Alternatively, a study that con-

tradicts the existing literature or for which there is a paucity of comparable studies should be considered with caution and the generalizability of its findings carefully explored.

The authors of the present study thoroughly discuss how their findings correlate with the existing literature and how this study might fill gaps in current knowledge. As with all research reports, readers should be cautious that selective reporting of literature that supports a study's findings is common, thus readers should take responsibility for conducting a broader search of the literature themselves.

Were All Clinically Important Outcomes Considered?

The effect of an intervention on any single outcome is typically only a small part of the relevant clinical picture. Particularly in studies of surgical interventions, important corollary outcomes invariably exist (e.g., postoperative complication rates, perioperative mortality, length of hospital stay, readmission rate, and quality of life, among others). Without this broader clinical picture, application of a study's results to a clinical setting is challenging and potentially dangerous.

For example, in the article by Kalata et al. [6], the authors concentrate mostly on bile duct injury. While this is an important complication to consider, the authors do not report on other important clinical outcomes such as perioperative mortality, readmission, and quality of recovery.

Are Limitations of the Study Clearly Described?

Authors should clearly describe the limitations of their study and how these limitations may impact the study's interpretation and application. Potential limitations in any study include those associated with study setting and study design. Issues such as selection bias, information/measurement bias, and measured and unmeasured confounding are present to some extent in every cohort study. The potential impact of these factors on the interpretation study results should be clearly described in the Discussion section, as exemplified by Kalata and colleagues [6] in the

paper under review. Specifically, the authors highlight the risk of information bias from relying on the accurate coding of variables in administrative databases. They also mention the unavailability of potentially important covariates such as sociodemographic factors, surgical complexity, and conversions to open that were not able to be measured. The authors describe how they attempted to mitigate these limitations, including the selection of reliably coded variables and sensitivity analyses including an instrumental variable analysis in an effort to control for unmeasured factors.

Resolution of the Clinical Scenario

The study by Kalata et al. [6] provides evidence calling into question the use of the robotic platform for minimally invasive cholecystectomy. The study design and the analytic approach used are appropriate for the clinical outcome of interest and the large, multi-institutional study population is likely generalizable to the population of interest. The authors demonstrate that robot-assisted cholecystectomy is associated with significantly greater rates of bile duct injury necessitating hepaticojejunostomy or choledocojejunostomy.

These results seem to be generalizable to the patient you are seeing in clinic. You discuss the results of this study with her at her next appointment, in the context of the study's limitations. After a thorough discussion, your patient decides to proceed with a conventional laparoscopic approach.

Conclusion

The article by Kalata et al. [6] exemplifies many of the strengths and a few of the weaknesses of cohort studies. The authors are mindful of many of the limitations of cohort studies and, where possible, take appropriate steps to mitigate them. The authors also take care to describe the current state of knowledge and how the present study fits into the broader clinical question.

In order for readers to meaningfully appraise research, reports must include sufficient detail. Reporting guidelines for several study designs have been created to standardize the presentation of research. The Strengthening The Reporting Of Cohort Studies in Surgery (STROCSS) guideline was developed in 2017 and updated in 2019, with the aim of improving the reporting quality of observational studies in surgery, including case-control, cross-sectional, and cohort studies [46]. Readers and investigators should be aware of these guidelines when appraising or conducting observational research.

While cohort studies cannot replace RCTs in terms of the quality of evidence, they play an important role in the surgical literature. Cohort studies are valuable in settings in which randomized trials may be unethical or unfeasible. Studies of rare diseases, studies of prognosis rather than therapy, and studies with extremely long-term outcomes are particularly well suited to a cohort-study design.

References

1. Exposure. In: Porta M, editor. A dictionary of epidemiology. Oxford: Oxford University Press; 2008:37.
2. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. *Evid Based Med*. 2016;21(4):125–7.
3. Laupacis A, Wells GA, Richardson WS, Tugwell P. Users' guides to the medical literature. V. How to use an article about prognosis. Evidence-Based Medicine Working Group. *JAMA*. 1994;272(3):234–7.
4. Schiphorst AH, Pronk A, Borel Rinkes IH, Hamaker ME. Representation of the elderly in trials of laparoscopic surgery for colorectal cancer. *Color Dis*. 2014;16(12):976–83.
5. Unger JM, Hershman DL, Albain KS, Moinpour CM, Petersen JA, Burg K, Crowley JJ. Patient income level and cancer clinical trial participation. *J Clin Oncol*. 2013;31(5):536–42.
6. Kalata S, Thumma JR, Norton EC, Dimick JB, Sheetz KH. Comparative safety of robotic-assisted vs laparoscopic cholecystectomy. *JAMA Surg*. 2023;158(12):1303–10.
7. Thoma A, Farrokhyar F, Bhandari M, Tandan V, Evidence-Based Surgery Working Group. Users' guide to the surgical literature. How to assess a randomized controlled trial in surgery. *Can J Surg*. 2004;47(3):200–8.
8. Cole P. The hypothesis generating machine. *Epidemiology*. 1993;4(3):271–3.
9. Loder E, Groves T, MacAuley D. Registration of observational studies. The next step towards research transparency. *BMJ*. 2010;340:c950. <https://doi.org/10.1136/bmj.c950>.
10. Should protocols for observational research be registered? *Lancet* 2010;375(9712):348.
11. Hernan MA. Causal analyses of existing databases: no power calculations required. *J Clin Epidemiol*. 2022;144:203–5.
12. Registries for evaluating patient outcomes: a user's guide. Rockville: Agency for Healthcare Research and Quality (US); 2014. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK208611/>. Accessed 20 Jul 2018.
13. Lawson EH, Louie R, Zingmond DS, Brook RH, Hall BL, Han L, Rapp M, Ko CY. A comparison of clinical registry versus administrative claims data for reporting of 30-day surgical complications. *Ann Surg*. 2012;256(6):973–81.
14. Cima RR, Lackore KA, Nehring SA, Cassivi SD, Donohue JH, Deschamps C, Vansuch M, Naessens JM. How best to measure surgical quality? Comparison of the Agency for Healthcare Research and Quality Patient Safety Indicators (AHRQ-PSI) and the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) postoperative adverse events at a single institution. *Surgery*. 2011;150(5):943–9.
15. Mason SA, Nathens AB, Byrne JP, Fowler R, Gonzalez A, Karanicolas PJ, Moineddin R, Jeschke MG. The accuracy of burn diagnosis codes in health administrative data: a validation study. *Burns*. 2017;43(2):258–64.
16. Lee DS, Stitt A, Wang X, Yu JS, Gurevich Y, Kingsbury KJ, Austin PC, Tu JV. Administrative hospitalization database validation of cardiac procedure codes. *Med Care*. 2013;51(4):e22–6.
17. Akl EA, Briel M, You JJ, Sun X, Johnston BC, Busse JW, Mulla S, Lamontagne F, Bassler D, Vera C, Alshurafa M, Katsios CM, Zhou Q, Cukierman-Yaffe T, Gangji A, Mills EJ, Walter SD, Cook DJ, Schünemann HJ, Altman DG, Guyatt GH. Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review. *BMJ*. 2012;344:e2809.
18. Shiloach M, Frencher SK Jr, Steeger JE, Rowell KS, Bartzokis K, Tomeh MG, Richards KE, Ko CY, Hall BL. Toward robust information: data quality and interrater reliability in the American College of Surgeons National Surgical Quality Improvement Program. *J Am Coll Surg*. 2010;210(1):6–16.
19. Bilimoria KY, Cohen ME, Ingraham AM, Bentrem DJ, Richards K, Hall BL, Ko CY. Effect of postdischarge morbidity and mortality on comparisons of hospital surgical quality. *Ann Surg*. 2010;252(1):183–90.
20. Dunn OJ. Multiple Comparisons Among Means. *J Am Stat Assoc*. 1961;56(293):52–64.
21. Armstrong RA. When to use the Bonferroni correction. *Ophthalmic Physiol Opt*. 2014 Sep;34(5):502–8.

22. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. 1990 Jan;1(1):43–6.
23. Sackett DL. Bias in analytic research. *J Chronic Dis*. 1979;32(1–2):51–63.
24. Delgado-Rodriguez M, Llorca J. Bias. *J Epidemiol Community Health*. 2004;58(8):635–41.
25. Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet*. 2002;359(9302):248–52.
26. Rothman KJ, Greenland S, Lash TL. *Modern epidemiology*. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008. 758 p.
27. Tripepi G, Jager KJ, Dekker FW, Wanner C, Zoccali C. Bias in clinical research. *Kidney Int*. 2008;73(2):148–53.
28. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–25.
29. Suissa S. Immortal time bias in pharmaco-epidemiology. *Am J Epidemiol*. 2008;167(4):492–9.
30. Gail MH. Does cardiac transplantation prolong life? A reassessment. *Ann Intern Med*. 1972;76(5):815–7.
31. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. *Nephron Clin Pract*. 2010;115(2):c94–9.
32. Greenland S, Morgenstern H. Confounding in health research. *Ann Rev Public Health*. 2001;22(1):189–212.
33. Mamdani M, Sykora K, Li P, Normand S-LT, Streiner DL, Austin PC, Rochon PA, Anderson GM. Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. *BMJ*. 2005;330(7497):960–2.
34. Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am J Epidemiol*. 2007;166(6):646–55.
35. Normand S-LT, Sykora K, Li P, Mamdani M, Rochon PA, Anderson GM. Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. *BMJ*. 2005;330(7498):1021–3.
36. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med*. 2010;29(20):2137–48.
37. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N Engl J Med*. 2007;357:2189–94.
38. Holodinsky JK, Austin PC, Williamson TS. An introduction to clustered data and multilevel analyses. *Fam Pract*. 2020;35(5):719–22.
39. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev Camb Philos Soc*, 5 ed. 2007;82(4):591–605.
40. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*. 1986;292(6522):746–50.
41. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14(5):365–76.
42. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf*. 2006;15(5):291–303.
43. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*. 2006;17(4):360–72.
44. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*. 1991;266(1):93–8.
45. Calder BJ, Phillips LW, Tybout AM. The concept of external validity. *J Consum Res*. 1982;9(3):240–4.
46. Mathew G, Agha R, for the STROCSS Group. STROCSS 2021: Strengthening the Reporting of cohort, cross-sectional and case-control studies in Surgery. *Int J Surg* 2021;96:106165.

Case-Control Studies

18

Achilles Thoma, Jenny Santos, Jessica Murphy,
Eric Duku, Emily Dunn, and Charles H. Goldsmith

Abbreviations

RCT	Randomized Controlled Trial
OR	Odds ratio
CI	Confidence Interval
EBM	Evidence-Based Medicine
GATE	Graphical Appraisal Tool for Epidemiological studies

Introduction

As explained in Chap. 5, the preferred method to determine the effects of a surgical intervention is a randomized controlled trial (RCT). However, while RCTs can minimize common research

biases, there are situations where performing an RCT is not feasible. Two such occasions include: (1) Ethical consideration of randomization and (2) The ability to detect the outcome of interest. Clinical trials in surgery aim to compare the effectiveness of novel procedures or interventions to the standard of care. A precondition to undertaking an RCT is the assumption that the novel intervention is not harmful, and we are in a state of “equipoise,” which means genuine uncertainty of the benefits of the novel intervention. Ethical issues arise when a patient is placed in potential danger; for example, we cannot ethically randomize a patient into a research arm that we suspect may cause harm or is potentially less effective than another option.

A. Thoma (✉)
Division of Plastic Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: athoma@mcmaster.ca

J. Santos
Southwestern Public Health (SWPH),
St. Thomas, ON, Canada

J. Murphy
Division of Physical Medicine & Rehabilitation,
Department of Medicine, McMaster University,
Hamilton, ON, Canada
e-mail: murphj11@mcmaster.ca

E. Duku
Department of Health Research Methods, Evidence,
and Impact, McMaster University,
Hamilton, ON, Canada

Offord Centre for Child Studies, Department
of Psychiatry and Behavioural Neurosciences,
McMaster University, Hamilton, ON, Canada
e-mail: duku@mcmaster.ca

E. Dunn
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: dunne4@mcmaster.ca

C. H. Goldsmith
Department of Health Research Methods, Evidence
and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: goldsmi@mcmaster.ca

Table 18.1 Comparison of cohort and case-control studies

	Cohort study	Case control study
<i>Advantages</i>	Gather data regarding the sequence of events, therefore can assess causality Examine multiple outcomes for an exposure Good for investigating rare exposures Can calculate rate of disease in exposed and unexposed	Good for rare outcomes or outcomes with long latency periods Relatively quick and inexpensive Require comparatively fewer patients Can utilize existing records Can examine multiple exposures/risk factors
<i>Disadvantages</i>	Large numbers of patients required Susceptible to selection bias Susceptible to recall bias or information bias ^a Less control over variables ^a May be expensive and require longer duration for follow-up ^b Susceptible to loss to follow-up ^b	Susceptible to recall or information biases Validation of information is difficult Possible incomplete ability to control for extraneous variables Possible difficulty selecting appropriate comparison group Rates of diseased in exposed and unexposed individuals cannot be determined
<i>Procedure</i>	1. Take disease-free (at risk) population 2. Identify exposed and unexposed cohort groups 3a. Identify diseased subjects by interview or written records ^a 3b. During follow-up identify the diseased subjects (incident cases) ^b 4. Analyze differences (i.e., incidence or relative risk) among the exposed and unexposed	1. Identify the cases (those who have the outcome of interest) 2. Select controls, which may be matched to cases 3. Measure exposure or risk factors of interest 4. Compare the presence or absence of exposure in the cases and controls
<i>Direction of Investigation</i>	^a From present to past ^b From Present to future	From present to past

Created with information from Song and Chung [2]

^aRetrospective Cohort Studies

^bProspective Cohort Studies

In terms of outcome detection, issues can occur if the outcome of interest is rare or takes a long period of time to develop. In these situations, an RCT would not be suitable as it would be both time-consuming and expensive, while the outcome may not occur within the timeframe of the study [1]. These issues leave clinicians with the choice of using a cohort or case-control study design.

While cohort and case-control studies are both forms of observational research they have individual differences, advantages, and disadvantages; these differences are summarized in Table 18.1. They are also explained in more detail in Chap. 17. Well-known examples of case-control studies include: (1) the 1926 study revealing that breast cancer risk increased with low fertility rates [2]; (2) the association between smoking and lung cancer [2]; and (3) the associa-

tion between the limb malformation, phocomelia, and thalidomide ingestion during pregnancy [1].

Clinical Scenario

At the pediatric surgical rounds, an uncommon case of chylothorax was presented in a three-month-old child who had heart surgery via a sternotomy incision. A previous heart surgery through a thoracotomy incision was unsuccessful. A discussion ensued as to the risk of chylothorax in this case. One pediatric cardiac surgeon claimed that this was due to the complexity of the congenital heart surgery performed whereas another claimed that this was due to the repeat surgery. The head of the service asks the pediatric cardiac fellow to review the literature on the association between chylothorax and repeat pediatric cardiac surgery.

Literature Search

As stated in the previous chapters, the first step to finding the best available evidence is to formulate a research question based on the PICOT format (See Chap. 3):

- **Population:** Pediatric Population
- **Intervention:** Cardiac surgery OR Redo Cardiac Surgery
- **Comparative intervention:** None
- **Outcomes:** Chylothorax
- **Time:** Any Time After Surgery

Using the PICOT format terms, your clinical research question could be one of two: “In paediatric patients undergoing cardiac surgery, what is the risk of chylothorax following surgery?” OR “In paediatric patients undergoing a redo cardiac surgery, what is the risk of chylothorax following surgery.”

Next, you use the above terms to perform a thorough literature search (see Chap. 4: *Finding the Evidence Through Searching the Literature*). Using the terms: “Paediatric” AND “Cardiac Surgery” AND “Chylothorax” you perform a literature search in the Cochrane Library to determine if there are any acceptable RCTs. The search yields two articles, however, neither are relevant to your research question (See Appendix A). You try performing another literature search on Cochrane using “Redo Cardiac Surgery” in place of “Cardiac Surgery” and this search yields no results. Since your search on Cochrane Library was not successful, you perform a literature search using PubMed using “Paediatric” AND “Cardiac Surgery” AND “Chylothorax.” Your search yields 14 publications so you decide to narrow your search to those published in between 2014 and February 1, 2024, which only removes 1 article, leaving 13 publications (See Appendix B). Twelve of these articles focused on the management of chylothorax or elements of this condition. One of these articles by Day et al. [3], titled “Chylothorax following paediatric cardiac surgery: A case-control study,” seemed promising. It is a matched case-control study investigating the association between numerous factors on the risk of chylothorax.

Appraisal of the Selected Article

When reviewing the article, you realize you are somewhat unfamiliar with the methodology and analysis used in case-control studies. As you want to be confident in the results that you will present to your colleagues, you search for a resource to help you better understand this study design. You find an article by Thoma et al. [1] that you feel can help you appraise this article before presenting it to your colleagues. The Thoma et al. [1] article provides appraisal questions specific to case-control studies where harm to a patient is involved; these questions are summarized in Box 18.1.

Box 18.1 Guidelines for the Appraisal of an Article in the Surgical Literature Using a Case-Control Study Design [1]

A. Are the results valid?

- Were cases and controls similar with respect to the indication or circumstances that would lead to exposure?
- Were the circumstances and methods for determining exposure similar for cases and controls?
- Was the correct temporal relationship demonstrated?
- Was there a dose–response relationship?

B. What were the results?

- How strong is the association between exposure and outcome?
- How precise was the estimate of risk?

C. Will the results help me in caring for my patients in my practice?

- Was the follow-up sufficiently long?
- Is the exposure similar to what might occur in my patient?
- What is the magnitude of risk?
- Are there any benefits known to be associated with the exposure?
- Were the patients in the appraised study similar to the patient in my practice?

Are the Results Valid?

- Were cases and controls similar with respect to the indication or circumstances that would lead to exposure?

The study by Day et al. [3] reviewed health records of all pediatric patients who received cardiac surgery at the Royal Children's Hospital (Australia). The authors used a 48-month window from January 2008 to January 2012, as the inclusion criteria for review. They authors identified 121 cases (diagnosed with chylothorax). The comparison (control) group was formed by identifying 121 patients who had also received heart surgery, within the same time period, but did not develop chylothorax. The control group was matched to the cases by: (1) Age (within 3 months if younger than 1 year of age or within 1 year for older patients); (2) Date of surgery (within 1 year); and (3) Sex. The matching of cases and controls is often performed in case-control studies to better control for confounding (the confusion of the effect of a risk factor on the outcome of interest, which is due to some other factor not accounted for in the study) as well as to provide a control group that better represents the target population. While, the authors controlled for these three factors, they did not explicitly check, with statistical testing, to ensure that the matching process worked for these variables.

The objective of the Day et al. [3] study was to explore the risk of developing chylothorax following surgery associated with a number of identified influencing factors. To explore this association, the cases and controls cannot be matched for these factors; the following factors have been identified as influencing the development of chylothorax following surgery and were explored by Day et al. [3]: (1) genetic risk factors; (2) the annual hospital volume [4]; (3) weight; (4) RACHS-1 score; (5) arch surgery; (6) open or closed heart; (7) incision site; and (8) surgery type (redo or virgin).

(ii) *Were the circumstances and methods for determining exposure similar for the cases and controls?*

When selecting cases, Day et al. [3] used the diagnostic criteria of: (1) a pleural fluid lympho-

cyte count of >80%, or (2) a triglyceride level > 1.1 mmol/L. Having predetermined diagnostic criteria helps to reduce misclassification bias, or the error associated with misidentifying a patient's disease status [1]. The authors then found a matched control group that had undergone cardiac surgery within the same time period but did not develop chylothorax. For both the case and control groups, the Cardiobase database from the Royal Children's Hospital in Melbourne was used.

(iii) *Was the correct temporal relationship established?*

Establishing a temporal relationship, meaning that the exposure occurred before the outcome, is important in determining whether there is an association. In the Day et al. [3] study, the authors do not give any information on the date of surgery as compared to the date of diagnosis. The absence of these data may likely be due to the fact that the objective of this study was to determine the diagnosis of chylothorax post-surgery, therefore, one would assume that the diagnosis date (outcome) would have come after the redo surgery date (exposure). While you understand why this information may not be provided, you feel it would have been helpful to be given the dates of surgery and the date of diagnosis. This information could have been beneficial to compare both the impact of the length of time between surgeries on risk, and the length of time past from surgery to diagnosis between the cases and controls.

(iv) *Was there a dose-response relationship?*

Determining a dose-response relationship can be difficult in surgery; it is much easier, for example, to find a dose-response relationship between smoking (number of packs per year) and lung cancer [1]. Previous work in the area of cardiac surgery and chylothorax have suggested a possible dose-response relationship between number of surgeries and the development of chylothorax. Both Ismail et al. [5] and Mery et al. [4] found that there was an increased incidence rate

in redo cardiac surgery cases. In the Day et al. [3] article, results indicate that there is an increased risk with redo surgery; therefore, this may be suggesting that there is a dose–response relationship regarding the number of interventions and chylothorax risk.

However, concluding a dose–response relationship, the study by Day et al. [3] may not be realistic as the authors did not use a measure detailing the specific number of redo surgeries in the analysis. The redo surgery variable was classified as “none” or “at least 1 redo.”

What Are the Results?

(i) *How strong is the association between exposure and outcome?*

In a case-control study, those individuals who have the outcome of interest are known before the study begins, therefore, a relative risk estimate cannot be used [1]. Instead, the odds ratio (OR) is more appropriate to measure the association of the exposure on the outcome [1]. An OR represents a ratio of the odds having the outcome in an “exposed” or case group relative to the odds of having the outcome in an “unexposed” or control group [6].

Day and colleagues matched the cases to the controls, and therefore a matched analysis method was used to calculate the OR as opposed to the standard method [1]. The standard approach for the OR is calculated using the number of individuals within both case and control groups who were or were not exposed to any risk factor(s) of interest for an unmatched study [1]. These numbers are best visualized by using 2X2 tables. This type of table is appropriate for unmatched studies because cases are independent from controls. Table 18.2 illustrates the basic format of a 2X2 table, to facilitate the calculation of an OR for every risk factor (exposure) included in an *unmatched* study. It is also important to note that calculating an OR by hand in this manner does not account for potential confounding by other factors (genetic risk factors, the annual hospital volume [4], weight at time of surgery, RACHS-1

Table 18.2 Basic format of a 2 × 2 table used in case-control studies

Exposure	Outcome	
	YES (Cases)	NO (Controls)
YES	a	b
NO	c	d

Table 18.3 Format of a 2X2 table that could have been used in the Day et al. [3] study had the design been *unmatched*

Surgery type	Chylothorax	
	Cases	Controls
REDO	36	9
VIRGIN	85	112

score, arch surgery, open or closed heart, and incision site [3]).

Using the format of this basic 2 × 2 table, an OR can be calculated using the following formula:

$$\text{Odds Ratio} = \frac{a/b}{c/d} = \frac{ad}{cb}$$

In the Day et al. [3] article, the exposure of interest was redo congenital cardiac surgery, while the outcome of interest was postoperative chylothorax. The authors included various risk factors for chylothorax following cardiac surgery including: the risk adjustment for congenital heart surgery (RACHS-1) score, arch surgery, open or closed heart, incision site and incision type (redo or virgin). A univariable OR is then calculated for each risk factor. For example, Table 18.3 illustrates what the 2X2 table for “Virgin surgery” versus “Redo surgery” would look like, using data from Day et al. [3], had the study been *unmatched* [3].

$$\begin{aligned} \text{Odds Ratio} &= \frac{ad}{bc} \\ &= \frac{(36)(112)}{(85)(9)} \\ &= 5.27 \end{aligned}$$

An OR of less than 1 represents a “protective” effect of the risk factor increasing the odds of the target group developing the target outcome [1]. In comparison, an OR of greater than 1 repre-

sents a “hazardous” effect [1]. For the above-mentioned scenario, you are interested in the OR associated with the repeat surgery. Using a multivariable conditional logistic regression model, Day et al. [3] reported an OR of 20.7 (95% Confidence Interval (CI): 4.24 to 100) for redo surgeries. This means that the odds of chylothorax developing in those exposed to a redo surgery is 20.7 times greater than those who are having a surgery for the first time after controlling for other confounding variables.

However, as cases and controls were matched by date of surgery, age at surgery, and sex, Day et al. [3] used a matched analysis, the basic format of the 2X2 table above would be slightly changed and the study results would have resulted from a different method of calculation. In a matched case-control, numbers of exposed and unexposed patients are illustrated by how many PAIRS had different exposure statuses (Table 18.4). Contrary to the traditional method, cases and controls are no longer independent groups and the unit of analysis is the matched case-control pair.

In turn, the formula shown above for an *unmatched* case-control study is also slightly changed to calculate the OR in a *matched* study:

$$\text{Odds Ratio} = \frac{b}{c}$$

In Table 18.4, the cells describe how many patients, within each *matched* pair (121 in total) were exposed or unexposed (redo or virgin surgery). For a matched case-control design, cells b and c are the only cells that would apply to calculate an OR, as you can be confident that the odds of having the outcome due to the exposure or not having the outcome due to not being exposed would be comparable between cases and controls in a matched sample.

In this situation it is not possible to calculate the OR that would represent the dependent nature

of cases and controls as the study does not provide information for matched pairs and just lists information by group for cases and controls. Again, calculating an OR in this manner does not account for potential confounding by other factors such as the risk factors listed previously.

Tables 18.2, 18.3, and 18.4 are purely for illustrative purposes, to show how you might go about getting a raw/unadjusted OR to measure the risk between the exposure and outcome in your study. This type of analysis is referred to as a univariable analysis. The univariable model in the Day et al. [3] article yielded an OR of 10.0, between redo surgeries and the risk of chylothorax. As this was significant at the 10% level in this analysis [3] (this is often used in univariable models to determine which risk factors may be confounding the results), it was one of the risk factors to be included in the subsequent multivariable model.

In this multivariable model, used to account for confounding by other factors, Day et al. [3] found an OR of 20.7, between redo surgeries and the risk of chylothorax. As the risk became greater in this model, which means that confounding factors influenced the relationship between chylothorax and “redo” surgeries, it was appropriate to include those factors in the multivariable model.

(ii) How precise was the estimate of risk?

The precision of an OR estimation can be judged by the Confidence Interval (CI), which is strongly impacted by the sample size of a study. Day et al. [3] used two methods of calculating the OR, a univariable conditional logistic regression model and a multivariable conditional logistic regression model. When looking at the univariable model, the “redo” group had an OR of 10.0 with a 95% CI of 3.05 to 32.8 [3]. In the multivariable model, the “redo” surgery group had an OR of 20.7 with a 95% confidence interval of 4.24 to 100 [3]. As a multivariable model accounts for other factors that can affect the relationship between an outcome and exposure, any conclusions should be based on the results from a multivariable model (OR = 20.7; 95% CI: 4.24 to 100).

Table 18.4 Format of a 2 × 2 table that should be used in the Day et al. [3] study due to its *matched* design

Case pair member	Control pair member	
	Exposed	Not Exposed
EXPOSED	a	b
NOT EXPOSED	c	d

The confidence intervals are quite wide, and most likely due to the low number of patients ($n = 45$) who had “redo” surgeries. While there are no set criteria to determine if a confidence interval is “too wide,” we can look at previous studies in the same area. Day et al. [3] state that their study is the second largest study in the area identifying 121 cases of chylothorax, second only to a multisite study [3]. Additionally, one should consider the fact that chylothorax is a relatively rare condition, therefore the absolute risk is quite low, and there is a smaller chance of this event occurring within a specific time interval. Therefore, while the estimate of the risk of chylothorax may not be overly precise, due to the small sample size, Day et al. [3] did have an adequate number of patients, as compared to other work in the area. Additionally, this shows that “redo” surgeries are quite rare in this study.

How Can I Apply the Results to My Patients in a Clinical Practice?

(i) *Was the follow-up sufficiently long?*

Determining if the length of follow-up was sufficient in a case-control study is different than in a prospective RCT or a cohort study [1]. As previously stated, the outcome of interest has already occurred in case-control studies before the onset of the investigation. Day et al. [3] included records in this study up to 48-months from January 2008, which is equivalent to approximately 4 years, or January 2012. A study period of 3 years seems sufficient for the outcome of interest; although some outcomes take time to develop, chylothorax has been found to develop on average, 8 days following surgery [8].

When comparing the current article to the literature, chylothorax studies seem to have a follow-up period varying from 2–7 years [7–10]. Further, Day et al. [3] identified 121 cases, making it the second largest study of its kind. Given the fact that the outcome has occurred prior to the commencement of the study, the similar follow-up interval to other studies, and the large number of cases identified, you are confident in concluding

that the follow-up in the Day et al. [3] study was sufficient.

(ii) *Is the exposure similar to what might occur in my patient?*

Where the exposure in the Day et al. [3] study is cardiac surgery, the exposure of interest in our scenario is redo heart surgery. Given that both exposures (cardiac surgery or redo heart surgery) are standardized procedures used for specific diseases or in specific situations, there is no reason to believe that there would be a difference between the exposure of your own patient, and the patients in the Day et al. [3] article.

(iii) *What is the magnitude of the risk?*

As previously explained, the odds ratio (OR) measures association between an exposure and an outcome. Additionally, the OR can also be used to compare the magnitude of the exposure for an outcome [10]. Box 18.2 summarizes how OR can demonstrate the effect of an exposure on an outcome.

Box 18.2 The Meaning of Different OR

OR = 1	The exposure does not affect the odds of the outcome occurring
OR > 1	The exposure is associated with higher odds of the outcome occurring
OR < 1	The exposure is associated with lower odds of the outcome occurring

Created using information from [10]

In the Day et al. [3] article, an OR of 20.7 was seen in the multivariable model, and an OR of 10.0 was seen in a univariate model. Both ORs demonstrate an increased risk of chylothorax with repeat cardiac surgery. In addition to the OR, confidence intervals can be used to measure the magnitude of association. While the confidence interval is used to determine precision, it can also identify the presence of statistical significance [10]. The 95% confidence interval can be used as a proxy for significance if the confi-

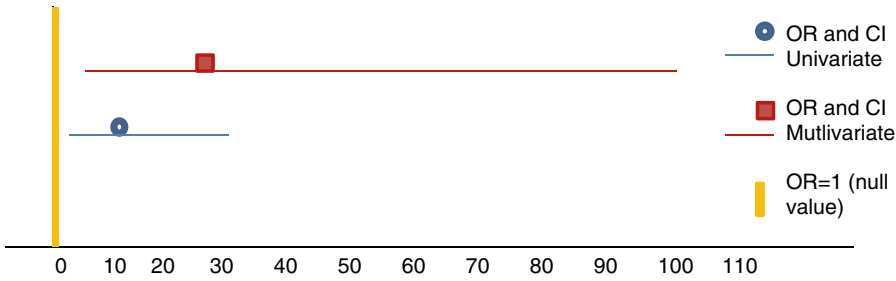


Fig. 18.1 Odds ratio and confidence interval for univariate and multivariable model

dence interval does not include 1, or the “null value” [10]. Figure 18.1 illustrates the OR and confidence interval for each model; in both cases, neither confidence interval includes the null value, implying significance.

overall risk for chylothorax seems quite low. Due to this low risk, the apparent commonality and mandatory nature of second surgeries you believe that there are benefits of this exposure to the patient.

(iv) *Are there any benefits that are known to be associated with the exposure?*

(v) *Were the patients in the appraised study similar to the patient in my practice?*

In this specific case, the “exposure” is a second surgery. Literature within cardiac surgery has discussed the common need for re-operation to address leftover or recurring defects from prior surgeries or to manage the breakdown or outgrowth of grafts [11]. In fact, an analysis of the Congenital Heart Surgery Database report that one-third of operations in the database were for reoperative cardiac surgery, for patients with congenital heart disease [11]. While it has been questioned over the years if reoperation is the cause for morbidity and operative mortality, studies over the past 20 years have concluded that there is minimal risk [11]. These minimal risks are due to enhancements in the quality of surgical techniques and perioperative care.

In the Day et al. [3] study, the median age for the case and control groups was 0.23 and 0.25 years, respectively. The patient in our scenario is 3 months old (or 0.25 years), the authors mention that those under 12 months of age are at higher risk and therefore, the results can likely be applied to our patient [3]. Additionally, this study was performed in Australia, which does not have any drastic differences in terms of healthcare, access to services, and lifestyle compared to the Canadian population [7]. Due to these similarities, you are confident that the information from the Day et al. [3] article could be generalized to your patient.

For chylothorax specifically, Day and colleagues [3] reviewed nine studies and reported that the incidence of chylothorax following a cardiac surgery varied from 0.85 to 15.8%. Day and colleagues [3] report an incidence rate of 5.23%. However, this is more of a prevalence rate; one cannot truly determine incidence from a case-control study as the outcome has already occurred before commencement and therefore includes both new and existing cases. However, we can still ascertain that the

Resolution of the Scenario

Based on the evidence presented in the Day et al. [13] article, the pediatric cardiac fellow presents her conclusions to her staff. Chylothorax is ten times as likely to occur with repeat cardiac surgery than a single previous surgery. She also acknowledges that the Day et al. [3] article also incriminates the complexity of the cardiac surgery especially around the arch of the aorta as carrying a higher risk for chylothorax. In this regard, both staff were correct.

**Additional Information
for the Reader**

Appraising epidemiological studies usually involves piecing together numerous components. The Graphical Appraisal Tool for Epidemiological studies (GATE) framework can help to organize these different elements of a study [12]. Figure 18.2 is our interpretation of this frame-

work, as it relates to the Day et al. [3] article. The original framework was developed by The Evidence-Based Medicine (EBM) Working group to help students conceptualize a study as a whole, as well as by its individual parts [4].

We have taken this framework and adjusted it to represent the variables that were looked at for this particular scenario. In our version of this framework, we used “Redo surgery” as the

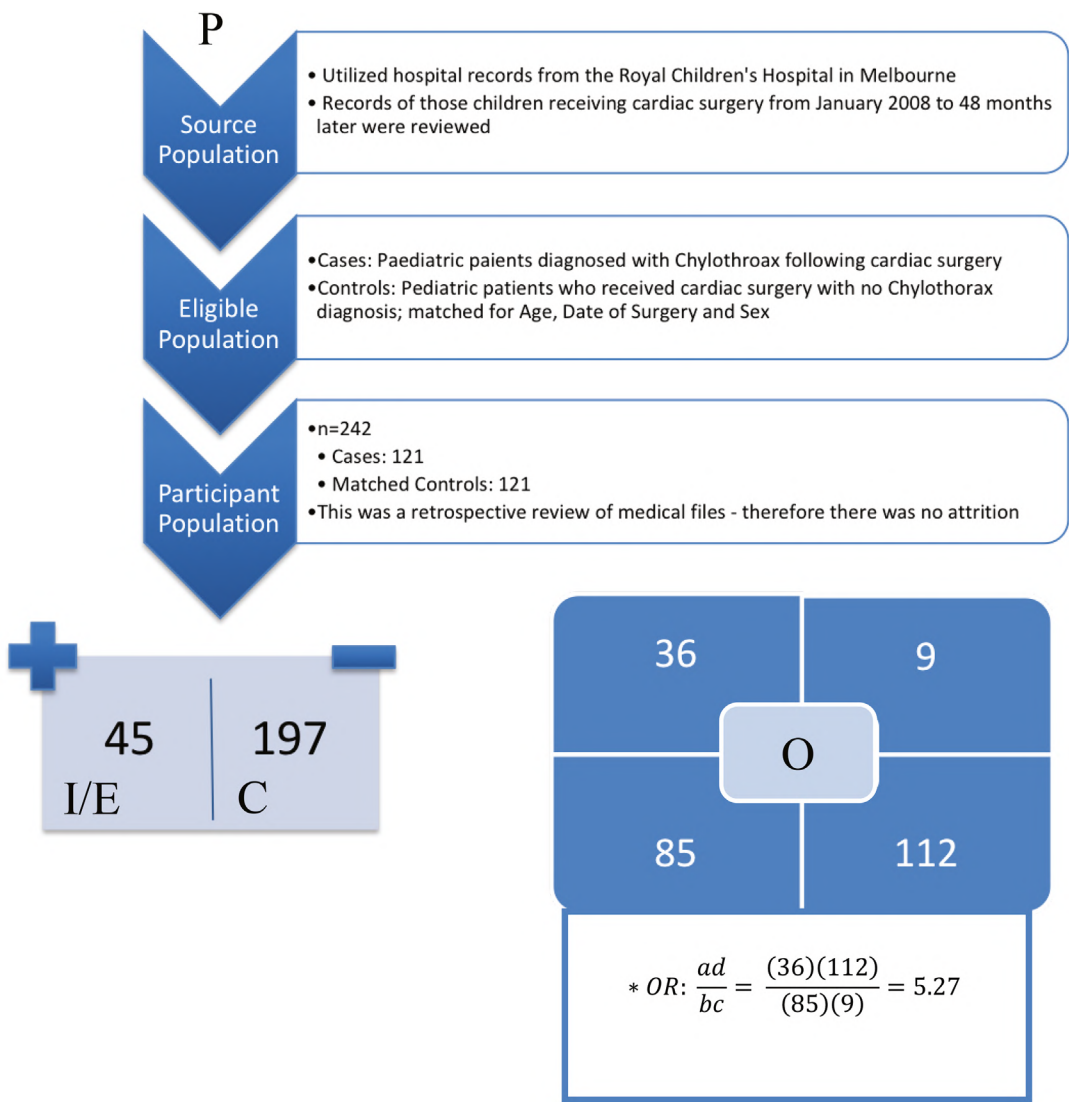


Fig. 18.2 GATE framework [12] interpretation for the Day et al. [3] article appraisal. (Note: P Population; I Intervention; E Exposure; C Comparison; OR

Odds Ratio. *; This calculation represents an OR using data from the Day et al. [3] study, had it been unmatched)

exposure/intervention and “Virgin surgery” as the comparison. We then explain the odds ratio calculation to show the risk associated with developing chylothorax (cases) following a “Redo surgery” (exposure). For this recreation of the framework, the time horizon was left out as there was no specific time horizon, other than just “following surgery” for this scenario. We encourage the reader to review the GATE framework as it offers a clear and concise format to summarize and appraise epidemiological studies [12].

The transparency of case-control studies is contingent not only on their methodological quality, which this chapter explores, but also on their reporting quality. For the clinical researchers who intend to undertake a case-control study we recommend they become familiar with the STROCSS 2021 guideline [13].

Appendix A

Search Results Using Cochrane Library: “Paediatric”, “Cardiac Surgery”, “Chylothorax”

1. Das A, Shah PS. Octreotide for the treatment of chylothorax in neonates. *Cochrane Database Syst Rev*. 2010;9:CD006388.
2. Mosalli R, AlFaleh K. Prophylactic surgical ligation of patent ductus arteriosus for prevention of mortality and morbidity in extremely low birth weight infants (Review). *Cochrane Database Syst Rev*. 2008;1:CD006181.

Appendix B

Search Results Using PubMed: “Paediatric,” “Cardiac Surgery,” “Chylothorax”

1. Day TG, Zannino D, Golshevsky D, d’Udekem Y, Brizard C, Cheyng MMH. Chylothorax following paediatric cardiac surgery: A case-control study. *Cardiol Young*. 2018;28(2):222–8.
2. Brown KL, Pagel C, Brimmell R, Bull K, Davis P, Franklin RC, Hoskote A, Khan N, Rodrigues W, Thorne S, Smith L, Chigaru L,

Utlely M, Wray J, Tsang V, Mclean A. Definition of important early morbidities related to paediatric cardiac surgery. *Cardiol Young*. 2017;27(4):747–56. <https://doi.org/10.1017/S1047951116001256>.

3. Buckley JR, Graham EM, Gaies M, Alten JA, Cooper DS, Costello JM, Domnina Y, Klugman D, Pasquali SK, Donohue JE, Zhang W, Scheurer MA. Clinical epidemiology and centre variation in chylothorax rates after cardiac surgery in children: a report from the Pediatric Cardiac Critical Care Consortium. *Cardiol Young*. 2017;1–8. <https://doi.org/10.1017/S104795111700097X>.
4. Brown KL, Pagel C, Ridout D, Wray J, Anderson D, Barron DJ, Cassidy J, Davis P, Hudson E, Jones A, Mclean A, Morris S, Rodrigues W, Sheehan K, Stoica S, Tibby SM, Witter T, Tsang VT; Cardiac Impact Study Group. What are the important morbidities associated with paediatric cardiac surgery? A mixed methods study. *BMJ Open*. 2019;9(9):e028533. <https://doi.org/10.1136/bmjopen-2018-028533>.
5. Lion RP, Winder MM, Amirnovin R, Fogg K, Bertrand R, Bhaskar P, Kasmai C, Holmes KW, Moza R, Vichayavilas P, Gordon EE, Trauth A, Horsley M, Frank DU, Stock A, Adamson G, Lyman A, Raymond T, Diaz I, DeMarco A, Prodhan P, Fundora M, Aljiffry A, Dewitt AG, Kozyak BW, Greiten L, Scahill C, Buckley J, Bailly DK; PC4 Chylothorax Work Group. Development of consensus recommendations for the management of post-operative chylothorax in paediatric CHD. *Cardiol Young*. 2022;32(8):1202–9. <https://doi.org/10.1017/S1047951122001871>.
6. Justice LB, Nelson DP, Palumbo J, Sawyer J, Patel MN, Byrnes JW. Efficacy and safety of recombinant tissue plasminogen activator for venous thrombosis after paediatric heart surgery. *Cardiol Young*. 2018;28(2):214–21. <https://doi.org/10.1017/S104795111700172X>.
7. Fogg KL, Trauth A, Horsley M, Vichayavilas P, Winder M, Bailly DK, Gordon EE. Nutritional management of postoperative chylothorax in children with CHD. *Cardiol Young*. 2023;33

- (9):1663–1671. <https://doi.org/10.1017/S1047951122003109>.
8. Itkin M, Chidekel A, Ryan KA, Rabinowitz D. Abnormal pulmonary lymphatic flow in patients with paediatric pulmonary lymphatic disorders: Diagnosis and treatment. *Paediatr Respir Rev*. 2020;36:15–24. <https://doi.org/10.1016/j.prrv.2020.07.001>.
 9. Adamson GT, Winder MM, Catton KG, Dewitt AG, Kozyak BW, Glenn ET, Bailly DK. Current practices for refractory chylothorax following congenital heart surgery. *Cardiol Young*. 2023;1–5. <https://doi.org/10.1017/S1047951123003918>.
 10. Haines C, Walsh B, Fletcher M, Davis PJ. Chylothorax development in infants and children in the UK. *Arch Dis Child*. 2014;99(8):724–30. <https://doi.org/10.1136/archdischild-2013-304364>.
 11. Winder MM, Eckhauser AW, Delgado-Corcoran C, Smout RJ, Marietta J, Bailly DK. A protocol to decrease postoperative chylous effusion duration in children. *Cardiol Young*. 2018;28 (6):816–25. <https://doi.org/10.1017/S1047951118000392>.
 12. Tashnizi MA, Maleki MH, Javedanfar O, Damsaz M, Alamdari AH, Seifalian AM, Asadi M, Hamidi Alamdari D. Platelet-rich plasma fibrin glue for treatment of chylothorax following cavopulmonary connections. *Eur J Cardiothorac Surg*. 2020;58(6): 1269–73. <https://doi.org/10.1093/ejcts/ezaa243>.
 13. Ootaki Y, Verghese GR, Ungerleider RM. Development of chylous ascites in an infant with ventricular septal defect: a case report. *Cardiol Young*. 2016;26(6):1225–7. <https://doi.org/10.1017/S1047951116000597>.
 3. Day TG, Zannino D, Golshevsky D, d’Udekem Y, Brizard C, Cheung MMH. Chylothorax following paediatric cardiac surgery: A case-control study. *Cardiol Young*. 2018;28(2):222–8.
 4. Mery CM, Moffett BS, Khan MS, Zhang W, Guzmán-Pruneda FA, Fraser CD Jr, Cabrera AG. Incidence and treatment of chylothorax after cardiac surgery in children: analysis of a large multi-institution database. *J Thorac Cardiovasc Surg*. 2014;147(3):678–86.
 5. Ismail SR, Kabbani MS, Najm HK, Shaath GA, Jijeh AMZ, Hijazi OM. Impact of chylothorax on the early post operative outcome after pediatric cardiac surgery. *J Saudi Heart Assoc*. 2014;26:87–92.
 6. Guyatt G, Rennie D, Meade MO, Cook DJ. Users guides to the medical literature: A manual for evidence-based clinical practice, 2nd ed. 2009 JAMA.
 7. Hussey P, Anderson GF, Osborn R, Feek C, McLaughlin V, Millar J, Epstein A. How does the quality of care compare in five countries? *Health Aff*. 2004;23(3):89–99.
 8. Yeh J, Brown ER, Kellogg KA, Donohue JE, Yu S, Gaies MG, Fifer CF, Hirsch JC, Aiyagari R. Utility of a clinical practice guideline in treatment of chylothorax in the postoperative congenital heart patient. *Ann Thorac Surg*. 2013;96:930–7.
 9. Chan EH, Russell JL, Williams WF, Van Arsdell GS, Coles JG, McCrindle BW. Postoperative chylothorax after cardi thoracic surgery in children. *Ann Thorac Surg*. 2005;80(5):1864–70.
 10. Szumilas M. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*. 2010;19(3):227–9.
 11. Villa-Hincapié CA, Careno-Jaimes M, Obando-Lopez CE, Camacho-Mackenzie J, Umaña-Mallarino JP, Sandoval-Reyes NF. Risk factors for mortality in reoperations for pediatric and congenital heart surgery in a developing country. *World J Pediatr Congenit Heart Surg*. 2017;8(4):425–9.
 12. Jackson R, Ameratunga S, Broad J, Connor J, Lethaby A, Robb G, Wells S, Glasziou P, Heneghan C. The GATE frame: critical appraisal with pictures. *Evid-Based Med*. 2006;11(2):35–8.
 13. Mathew G, Agha R, Albrecht J, Goel P, Mukherjee I, Pai P, D’Cruz AK, Nixon IJ, Roberto K, Enam SA, Basu S, Muensterer OJ, Giordano S, Pagano D, Machado-Aranda D, Bradley PJ, Bashashati M, Thoma A, Afifi RY, Johnston M, Challacombe B, Ngu JC, Chalkoo M, Raveendran K, Hoffman JR, Kirshtein B, Lau WY, Thorat MA, Miguel D, Beamish AJ, Roy G, Healy D, Ather HM, Raja SG, Mei Z, Manning TG, Kasivisvanathan V, Rivas JG, Coppola R, Ekser B, Karanth VL, Kadioglu H, Valmasoni M, Noureldin A; STROCSS Group. STROCSS 2021: Strengthening the reporting of cohort, cross-sectional and case-control studies in surgery. *Int J Surg*. 2021;96:106165.

References

1. Thoma A, Kaur MN, Farrokhyar F, Waltho D, Levis C, Lovrics P, Goldsmith CH. Users’ guide to the surgical literature: How to assess an article about harm. *Can J Surg*. 2016;59(5):351–7.
2. Song JW, Chung KC. Observational studies: Cohort and case-control studies. *PRS*. 2010;126(5):2234–42.

Evaluating Case Series in Surgery

19

Christopher J. Coroneos and Brian Chin

Abbreviations

BI-ALCL	Breast Implant-associated Anaplastic Large Cell Lymphoma
LTFU	Loss to follow-up
RCT	Randomized controlled trial

Introduction

Case series is an observational study design involving a group of similar patients, who share a common exposure, diagnosis, intervention, or outcome [1]. In surgery, they most commonly report rare/unique cases, or the results of novel interventions [2]. Several landmark disease classifications and interventions were first reported in case series [1]. Given the focus on technique and innovation in surgical literature, it is not surprising that most top historical citations, and even current publications,

are case series/case reports [1, 3–5]. From a methodological perspective, case series represent low-level evidence (Level IV) [6]. With the absence of a control group, and commonly retrospective design, they are especially prone to bias [6]. However, case series remains a useful design in surgical research for generating hypotheses and feasibility of interventions [7]; patient characteristics and outcomes are often used to plan higher level studies (e.g., sample size calculation, follow-up timing) [6, 7]. A case series alone is often sufficient to establish diagnostic accuracy or demonstrate safety [6]. Case series can demonstrate “real-world” effectiveness, with high external validity [7, 8]. By design, they remain cheaper and faster to complete than higher levels of evidence [2]. Given their prevalence in surgical literature, it is important every surgeon be familiar with the critical appraisal of case series. Critical appraisal should follow the same principles as with a cohort study, with attention to bias introduced by lack of a control group. The key to their assessment is attention to selection bias, recall bias, and information bias.

C. J. Coroneos (✉)

Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada

Department of Health Research Methods, Evidence,
and Impact, McMaster University, Hamilton,
ON, Canada

e-mail: coronec@mcmaster.ca

B. Chin

Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada

e-mail: chinh@mcmaster.ca

Clinical Scenario

You are a plastic surgeon working at a cancer center. Your reconstructive practice includes lymphedema. Your colleagues in breast surgery request you see a patient with upper extremity lymphedema following previous breast cancer

treatment. She has previously undergone lymphovenous bypass, but now has advanced lymphedema. She is indicated for a lymph node transfer but would prefer to not use her contralateral axilla or groins as donor sites given the risk of iatrogenic lymphedema in another limb. Having recently attended a microsurgical conference, you recall a presentation on vascularized lymph node transfer using mesenteric lymph nodes, where lymph nodes from the jejunal mesenteric root are used as a donor. You wonder if the surgeons that presented the work have published their technique in a peer-reviewed journal.

Literature Search

You visit [PubMed.gov](#) and perform a search with the terms: “mesenteric vascular lymph node lymphedema.” Your search is limited to

English, without date restrictions ([Appendix 1](#)). This yields 13 results ([Appendix 1](#)); on review of abstracts on this list, references [3, 6] are review articles, 5 and 8 are anatomic studies, 10 and 12 focus on gynecologic oncology, 11 and 13 are basic science studies, 7 and 9 are letters outlining a procedure. Reference [2] reports outcomes of a variety of lymph node transfers, and 4 focused on complications. You select a reference that has a recent case series published in *Plastic and Reconstructive Surgery* by Schaverien et al. [9] from MD Anderson Cancer Center (Houston, Texas) including 25 patients. The key characteristics of the case series are listed in Table 19.1.

Table 19.1 Key characteristics of the case series by Schaverien et al. [9]

Characteristic	Case series
Objective	“...report outcomes of our modified jejunal mesenteric vascularized lymph node trans-plantation technique in patients treated for upper extremity lymphedema.”
Population	n: 25 Age (mean, SD): 55.9 (10.1) years <i>MDACC Staging</i> III: 15 IV: 10 Cellulitis History: 11 <i>Cancer</i> Breast cancer: 24 Melanoma: 1 <i>Surgical</i> Mastectomy: 20 Breast conservation: 3 Axillary lymph node dissection: 25 <i>Adjuvant Therapy</i> Radiation therapy: 24 Chemotherapy: 23 Active smoking: 0 Quit smoking (>4 weeks): 7 Hypertension: 3 Diabetes mellitus: 1
Intervention	Single-level transfer: 10 Dual-level transfer: 15 Upper midline laparotomy: 8 Transverse: 17

(continued)

Table 19.1 (continued)

Characteristic	Case series
Study design	Retrospective review of prospective database Consecutive patients
Outcomes	Limb volume (Perometer) Bioimpedence spectroscopy (LDex) Lymphedema Life Impact Scale Quick disabilities of the arm, shoulder, and hand Cellulitis
Follow-up	13.4 ± 6.8 months
Complications	Flap related: 0 Ischemic bowel: 0

Appraisal of a Surgical Case Series

The appraisal of case series involves evaluating the validity of the study, interpreting the results, and applying study findings clinically (Box 19.1).

Box 19.1 Framework for the Appraisal of Surgical Case-Series

I. *Are the results valid?*

- (i) Is a clear objective stated?
- (ii) Is the series a prospective study?
If not, was data collection prospective?
- (iii) Was patient recruitment consecutive?
- (iv) Was outcome assessment appropriate?
- (v) Is the intervention appropriately described?

II. *What are the results?*

- (i) Are outcomes appropriately reported?
- (ii) Are outcomes completely reported?

III. *Will the results change practice?*

- (i) Are the patients similar to my own?
- (ii) Is the setting similar to my own?

I. *Are the Results Valid?*

(i) *Is a clear objective stated?*

The case series should have a clear objective, as with any other study design; it can be reported as a stated aim, or more appropriately as a structured research question. The common patient factors (e.g., diagnosis, intervention) and rationale for reporting patients should be defined. The objective should be pertinent and add to the current body of research on the topic. The series by Schaverien et al. [9] states an objective to “*report outcomes of our modified jejunal mesenteric vascularized lymph node transplantation technique in patients treated for upper extremity lymphedema*.” There is no structured research question stated. The statement defines the general population and intervention but does not incorporate the technical modifications that are the focus of the paper in optimizing outcomes. These aspects are alluded to in the title, or subsequently discussed in the manuscript.

(ii) *Is the series a prospective study? If not, was data collection prospective?*

Prospective study designs demonstrate more rigorous methodology; patient inclusion factors are predefined, baseline characteristics are accurately captured, the intervention is consistent, and outcomes and their timing are established a

priori. Patients can be captured at the same time point in their disease process, ideally at initial presentation. This is especially important in studies of exposure; temporal relationships between exposure and acute versus long-term outcomes can be established [10]. It is often termed an “inception cohort” and should not be confused with a cohort study. As corollary, readers must be aware of information bias, where there is excessive probing into risk factors/outcomes biasing results; data are not collected in a standard manner [11]. Prospective designs are unfortunately more expensive and labor-intensive [10].

If the entire case series does not follow a prospective protocol, data collection could have been prospective. Data quality is optimized if established a priori [10], with uniform fields for each patient. A common alternative strategy for surgeons is to use their existing prospective patient database if one exists. Retrospective assessment of clinical and administrative records may result in missing outcomes of interest and inconsistent data. Retrospective patient assessment is prone to recall bias, where remembered risk factors or outcomes are inaccurate.

Schaverien et al. [9] report that the series has a retrospective design; however, they use a prospectively collected database. Given that the series reports results of a surgical intervention, the reconstructive procedure itself is a consistent time point of reference for all patients. Further, the patients are homogeneous in terms of their presentation, previous disease status, and treatment to date. While a prospective study would be ideal, recall and information bias are minimized in this series.

(iii) *Was patient recruitment consecutive?*

Inclusion of consecutive patients is the best indication to the reader that selection bias is minimized. In case series where favorable patients are selected or “cherry-picked,” results will likely overestimate positive outcomes, namely, the success of a surgical intervention [7]. For example, selecting only low-risk patients will result in decreased morbidity and mortality. Reporting only ideal patients limits the external validity of

the study, and readers cannot be confident the series represents practical effectiveness for all patients in real-world practice [12]. The inclusion of consecutive patients is the best approximation of the treatment group in a cohort study. For case series reporting a high-risk surgical intervention, consecutive cases undergoing surgery do not represent all patients with the disease; some discussion of the characteristics for patients who agree to treatment should be included. In case series where consecutive patients are not included, specific inclusion and exclusion criteria should be reported [7]; readers should interpret results accordingly.

In a special case, for series where an incidence or estimate of risk is reported, it is important that all patients are reported for a given time period and represent a demographic patient sample. An accurate “denominator” for the patient group must be obtained. For example, a series of all births at a hospital are a demographic sample, and the study can provide the incidence of birth injury for the given patient group. Conversely, a series of cases not resolving with conservative management and subsequently referred to a specialist represents only a portion of all cases, and patient group metrics cannot be calculated.

Schaverien et al. [9] specifically report consecutive patients with specific inclusion and exclusion criteria. The series reports that patients underwent surgery from May 2017 to February 2020 with a jejunal mesenteric vascularized lymph node transplantation to treat unilateral upper extremity lymphedema at MD Anderson Cancer Center (Houston, Texas). This series should represent their entire experience with the procedure.

(iv) *Was outcome assessment appropriate?*

As with other study designs, the primary outcome of a case series should be specific, clinically relevant, measured at an appropriate time, and be important to patients; these factors are indicative of the series’ clinical impact [13]. Readers can consider the modified hierarchy of patient-important outcomes where Class I: mortality, Class II: morbidity, Class III: symptoms/

quality of life/functional status, Class IV: surrogate [14, 15]. The highest-class clinically relevant outcome should be reported to have the largest impact on patient care [14, 15]. Class I and II outcomes are considered “hard” outcomes; they are dichotomous, and least subjective (e.g., “life or limb”). For example, a case series on melanoma, or major burn care, can report mortality (Class I), a case series on early laparoscopic cholecystectomy for cholecystitis can measure 30-day major complications (Class II), and a case series on breast reconstruction can report quality of life with BREAST-Q [16] (Class III).

Appropriate timing of outcome assessment is equally important; the reader must consider if sufficient time has accumulated to assess patient survival, the success of a procedure, or quality of life [7]. For example, 30 days may be appropriate for a series assessing survival of acute trauma interventions, whereas 2 years may be necessary for long-term functional outcomes of a nerve reconstruction potentially requiring multiple procedures or long-term physical therapy. Readers should look for a minimum follow-up period as a criterion for case series inclusion [6]. Finally, possible bias in outcome assessment should be considered. Blinding outcome assessment is frequently feasible in prospective designs [13]. For retrospective designs, hard outcomes are the most objective.

Schaverien et al. [9] report several appropriate postoperative outcomes. Class II outcomes of flap loss, length of hospital stay, and cellulitis are reported. The study does not report measures of donor site morbidity. Class III outcomes of lymphedema-specific and general patient-reported outcome measures are reported at fixed intervals. Class IV outcomes include volume assessment by Perometer volumetry and bioimpedance spectroscopy.

(v) *Is the intervention appropriately described?*

Most surgical case series describe a procedure; technical details of the intervention and perioperative management should be sufficiently described for readers to appropriately interpret, and possibly reproduce reported results [7]. Beyond the specific description of a surgical pro-

cedure, readers should evaluate for operative indications, preoperative work-up/care (e.g., imaging investigations, resuscitation protocols), perioperative care (e.g., antibiotics, venous thromboembolism prophylaxis, necessary instruments/disposables, dosages of unique medications), and postoperative care (e.g., physical therapy protocols, imaging surveillance) [1]. Any co-interventions should be described [17], (e.g., comorbid fracture repair with intramedullary nail of lower extremity fractures, abdominoplasty with breast reduction) [7].

Schaverien et al. [9] include a section titled *Patient Selection for Surgical Intervention* outlining the authors’ baseline evaluation, imaging, indications, and preoperative optimization for this procedure in a transparent manner. The included population is discussed in a nuanced manner, with special consideration to specific patient presentations, as to which mesenteric nodes are best suited for, and relative contraindications to the procedure. Readers should understand that the reported outcomes of the procedure can only be applied to patients meeting the same criteria.

Authors comprehensively report the details of the flap harvest, donor site preparation, and in setting with a separate section titled *Surgical Technique*. The procedure is meticulously described, with added details for visualizing ideal nodes and management of the recipient site with accompanying videos. Beyond the meticulous description of the intervention, technical refinements, pitfalls, and the “how” and “why” of these critical steps are rationalized. Authors report 17 patients underwent co-intervention with lymphovenous bypass.

II. *What are the Results?*

(i) *Are outcomes appropriately reported?*

The case series is a descriptive study design, not analytic; no control group is present [6]. As such, “cause and effect” relationships cannot be established. Readers should be cautious that causal inferences between an intervention and outcomes are not reported. Instead, outcomes can be discussed in a

descriptive manner, and hypotheses can be generated for testing in a higher-level study design [6]. For example, the first case series of breast implant-associated anaplastic large cell lymphomas (BI-ALCL) [18] could not demonstrate causation of silicone breast implants; results suggested an association with silicone implants, and suggested confirmation with higher-level studies (e.g., case-control study). As discussed in *Was patient recruitment consecutive?* the series also did not include a demographic patient sample, and incidence of BI-ALCL could not be estimated.

Authors should be conservative in reporting case series results [6]. In select cases, authors may report analyses “before and after” a procedure or compared to historical controls. Any such analyses should have been defined a priori and have rational sample size calculations. Special attention should be directed to repeated measures over time. Patients receiving novel or investigational procedures are likely distinct from the general population [17]. Historical controls are rarely an adequate comparison group; patients treated previously were in completely distinct settings, and analyses are thus prone to several biases [17].

Schaverien et al. [9] report essential outcomes (as discussed in *Was outcome assessment appropriate?*). Most analyses are appropriate. Some may benefit from different statistical approaches. For example, authors report several repeated outcome measurements with continuous variables at 3-, 6-, and 12-month intervals. These appear to have been analyzed with t and Mann-Whitney U tests. In a case series, a more appropriate choice for repeated measures for one group of patients would be a repeated measures ANOVA, to better incorporate changes over time.

(ii) *Are outcomes completely reported?*

Complete outcomes, especially beyond those demonstrating benefit, should be reported in a case series. In a prospective series, any loss to follow-up (LTFU) should be reported. Readers

can assume patients LTFU have outcomes worse than those completing the study [19]. Patients with poor outcomes or health states do not present for follow-up, and these issues are not accurately captured (e.g., patients with cancer not presenting for follow-up with their reconstructive surgeon may mean they are deceased). Further, reporting LTFU is a sign to the reader that a study was completed with higher methodological rigor [20]. For those identified LTFU, some discussion of the tolerability of the intervention, or treatment attrition is appropriate. In retrospective series, readers can similarly assess for missing outcome data; fewer missing data is a sign that a robust database was used.

In series where mortality or morbidity are not the most clinically important outcomes, reporting harms and potential complications is necessary. Readers should look for conventional reporting of overall major complications, re-intervention, readmission, and appropriate intra-operative harms (e.g., blood loss, damage to surrounding structures) [1, 7]. In case series including patients treated over a long period of time, readers should be critical of identifiable and possible unknown changes to the patient population, disease characteristics, and intervention [6]; changes to the surgical procedure over time should be outlined, and those changes should be rationalized to the reader. Similarly, authors can indicate the presence of a learning curve to the reader [1, 7].

Schaverien et al. [9] report mean follow-up time of 13.4 ± 6.8 months was achieved, implying not all patients reached 12-month reported follow-up for postoperative outcomes. It is assumed, but not stated, that all patients reached 6-month follow-up based on described postoperative management. When discussing 12-month cellulitis, patient outcomes, and discontinuation of compression therapy, it is not clear how many patients reached this length of follow-up. Authors do not comment on a learning curve for this procedure. Instead, they note mesenteric flaps are “harvested by a plastic surgeon with expertise in raising jejunal flaps for esophageal reconstruction.” A surgeon without this previous experience would likely experience a learning curve.

III. *Will the Results Change Practice?*

(i) *Are the patients similar to my own?*

After considering the validity of the case series, and appraising the results, readers must decide if patients are comparable to their own. Case series typically have high external validity [7]. Inclusion and exclusion criteria are not as stringent as in higher-level study designs (e.g., randomized trials), and results can often be applied to a wider spectrum of patients [6]. Still, readers should consider the comorbidities and characteristics of the patients included in the case series, especially prognostic variables (e.g., age, smoking status, tumor stage) [1].

Schaverien et al. [9] comprehensively describe their patient population. Table 19.1 describes baseline demographics, with patients typically presenting following breast cancer treated with mastectomy, axillary lymph node dissection, radiation, and chemotherapy. These prognostic factors are however not surprising given the disease process and intervention described in the case series.

(ii) *Is the setting similar to my own?*

In applying results, readers must be certain that the care setting is similar to their own. Case series are often single center; however, if outcomes are pooled with a multicenter design, the differences between centers should be defined [7]. Novel and innovative interventions are often introduced by tertiary or even “quaternary” settings, where highly specialized care is delivered to a worldwide referral base. While academic centers typically have more resources versus community settings, patients are correspondingly more complex [7]. Multidisciplinary care is often necessary for increasingly complex interventions, for example, intraoperative anesthesia management, expertise of other surgical specialties, postoperative inpatient monitoring, postoperative imaging surveillance and functional assessment, and long-term physical therapy.

Schaverien et al. [9] report a series from MD Anderson Cancer Center, which can be classified

as a “quaternary” setting, where patients are captured from a global base [21]. Expertise beyond a typical microsurgeon is necessary, with the series describing “*a plastic surgeon with expertise in raising jejunal flaps for esophageal reconstruction*,” and co-authors including a thoracic surgeon. These surgeons support an advanced reconstructive technique. Further, co-authors include a specialized physician assistant, and a Certified Lymphedema Therapist. This multidisciplinary team is likely necessary to perform preoperative assessments, “prehabilitation” with optimized conservative therapy, and consistent postoperative assessment.

Conclusion

Although case series are recognized as low Level IV in the hierarchy of evidence, they remain prevalent in surgical literature. Case series are the most common study design in surgical research, and remain the predominant study design in reporting unique patient presentations, and innovative surgical procedures. They are unique in their high external validity and fast and less costly design. Alone, case series can establish the safety of an intervention, or the diagnostic accuracy of a test [6]. From a methodological standpoint, they are important in demonstrating feasibility, and patient characteristics for higher-level study planning. Appraising a case series is similar to assessing the methodology and treatment group of a cohort trial. With consideration of the topics discussed in this chapter, surgeons can confidently assess the validity and applicability of a case series relevant to their practice.

Appendix 1—Search Strategy

<https://pubmed.ncbi.nlm.nih.gov/?term=mesenteric+vascular+lymph+node+lymphedema&filter=lang.english>

Pubmed String: mesenteric vascular lymph node lymphedema

Article Language: English

Publication Date: all

Results:

1. Schaverien MV, Hofstetter WL, Hall MS, Chen DN, Selber JC. Jejunal mesenteric vascularized lymph node transplantation for lymphedema: Outcomes and technical modifications. *Plast Reconstr Surg*. 2022;149(4):700e–10e. <https://doi.org/10.1097/PRS.0000000000008960>. PMID: 35157612.
2. Schaverien MV, Asaad M, Selber JC, Liu J, Chen DN, Hall MS, Butler CE. Outcomes of vascularized lymph node transplantation for treatment of lymphedema. *J Am Coll Surg*. 2021;232(6):982–94. <https://doi.org/10.1016/j.jamcollsurg.2021.03.002>. Epub 2021 Mar 22. PMID: 33766726.
3. Chu CK, Schaverien MV, Chang EI, Hanson SE, Hanasono MM, Selber JC. Intra-abdominal lymph nodes: A privileged donor site for vascularized lymph node transfer. *Plast Reconstr Surg Glob Open*. 2020;8(3):e2673. <https://doi.org/10.1097/GOX.0000000000002673>. PMID: 32537337; PMCID: PMC7253245.
4. Kraft CT, Eiferman D, Jordan S, Skoracki RJ. Complications after vascularized jejunal mesenteric lymph node transfer: A 3-year experience. *Microsurgery*. 2019;39(6):497–501. <https://doi.org/10.1002/micr.30491>. Epub 2019 Jul 8. PMID: 31283856.
5. Ruter D, Chen W, Garza R 3rd, Eiferman D, Skoracki R. Mesoappendix as potential donor site for vascularized lymph node transfer: anatomic study. *J Surg Res*. 2018;230:143–7. <https://doi.org/10.1016/j.jss.2018.03.041>. Epub 2018 May 25. PMID: 30100031.
6. Schaverien MV, Badash I, Patel KM, Selber JC, Cheng MH. Vascularized lymph node transfer for lymphedema. *Semin Plast Surg*. 2018;32(1):28–35. <https://doi.org/10.1055/s-0038-1632401>. Epub 2018 Apr 9. PMID: 29636651; PMCID: PMC5891655.
7. Schaverien MV, Hofstetter WL, Selber JC. Vascularized jejunal mesenteric lymph node transfer for lymphedema: A novel approach. *Plast Reconstr Surg*. 2018;141(3):468e–9e. <https://doi.org/10.1097/PRS.0000000000004162>. PMID: 29485606.
8. Coriddi M, Wee C, Meyerson J, Eiferman D, Skoracki R. Vascularized jejunal mesenteric lymph node transfer: a novel surgical treatment for extremity lymphedema. *J Am Coll Surg*. 2017;225(5):650–7. <https://doi.org/10.1016/j.jamcollsurg.2017.08.001>. Epub 2017 Aug 14. PMID: 28818700.
9. Coriddi M, Skoracki R, Eiferman D. Vascularized jejunal mesenteric lymph node transfer for treatment of extremity lymphedema. *Microsurgery*. 2017;37(2):177–178. <https://doi.org/10.1002/micr.30037>. Epub 2016 Feb 18. PMID: 26892278.
10. Narducci F, Lambaudie E, Mautone D, Hudry D, Bresson L, Leblanc E. Extraperitoneal para-aortic lymphadenectomy by robot-assisted laparoscopy in gynecologic oncology: preliminary experience and advantages and limitations. *Int J Gynecol Cancer*. 2015;25(8):1494–502. <https://doi.org/10.1097/IGC.0000000000000504>. PMID: 26270116.
11. Liu X, Pasula S, Song H, Tessneer KL, Dong Y, Hahn S, Yago T, Brophy ML, Chang B, Cai X, Wu H, McManus J, Ichise H, Georgescu C, Wren JD, Griffin C, Xia L, Srinivasan RS, Chen H. Temporal and spatial regulation of epsin abundance and VEGFR3 signaling are required for lymphatic valve formation and function. *Sci Signal*. 2014;7(347):ra97. <https://doi.org/10.1126/scisignal.2005413>. PMID: 25314967; PMCID: PMC4226761.
12. Ohba Y, Todo Y, Akashi D, Kawaguchi S, Minobe S, Okamoto K, Kato H. Incidence of metastasis in the circumflex iliac nodes distal to the external iliac nodes in ovarian cancer. *Gynecol Oncol*. 2012;126(3):387–90. <https://doi.org/10.1016/j.ygyno.2012.05.031>. Epub 2012 Jun 1. PMID: 22664060.
13. Galanzha EI, Tuchin VV, Zharov VP. Optical monitoring of microlymphatic disturbances during experimental lymphedema. *Lymphat Res Biol*. 2007;5(1):11–27. <https://doi.org/10.1089/lrb.2007.5103>. PMID: 17508899.

References

1. Agha RA, Fowler AJ, Rajmohan S, Barai I, Orgill DP, PROCESS Group. Preferred reporting of case series in surgery; the PROCESS guidelines. *Int J Surg*. 2016;36:319–23.
2. Chan K, Bhandari M. Three-minute critical appraisal of a case series article. *Indian J Orthop*. 2011;45:103–4.
3. Joyce KM, Joyce CW, Kelly JC, Kelly JL, Carroll SM. Levels of evidence in the plastic surgery literature: a citation analysis of the top 50 'classic' papers. *Arch Plast Surg*. 2015;42:411–8.
4. Sochacki KR, Jack RA, Nauert R, Harris JD. Correlation between quality of evidence and number of citations in top 50 cited articles in rotator cuff repair surgery. *Orthop J Sports Med*. 2018;6:232596711877663.
5. Cordero JJ, Eidelson S, Frediani T, Shakoori P, Carré AL, Klausmeyer MA, Chu MW. The top 100 cited articles in the microsurgical treatment for lymphedema. *J Reconstr Microsurg*. 2023;39:559–64.
6. Kooistra B, Dijkman B, Einhorn TA, Bhandari M. How to design a good case series. *J Bone Joint Surg Am*. 2009;91(Suppl 3):21–6.
7. Coroneos CJ, Ignacy TA, Thoma A. Designing and reporting case series in plastic surgery. *Plast Reconstr Surg*. 2011;128:361e–8e.
8. Audigé L, Hanson B, Kopjar B. Issues in the planning and conduct of non-randomised studies. *Injury*. 2006;37:340–8.
9. Schaverien MV, Hofstetter WL, Hall MS, Chen DN, Selber JC. Jejunal mesenteric vascularized lymph node transplantation for lymphedema: outcomes and technical modifications. *Plast Reconstr Surg*. 2022;149:700e–10e.
10. Bernard T, Armstrong-Wells J, Goldenberg N. The institution-based prospective inception cohort study: design, implementation, and quality assurance in pediatric thrombosis and stroke research. *Semin Thromb Hemost*. 2012;39:010–4.
11. Sedgwick P. Prospective cohort studies: advantages and disadvantages. *BMJ*. 2013;347:f6726.
12. Sedgwick P. Bias in observational study designs: prospective cohort studies. *BMJ*. 2014;349:g7731.
13. Voineskos SH, Coroneos CJ, Ziolkowski NI, Kaur MN, Banfield L, Meade MO, Thoma A, Chung KC, Bhandari M. A systematic review of surgical randomized controlled trials: Part I. Risk of bias and outcomes: Common pitfalls plastic surgeons can overcome. *Plast Reconstr Surg*. 2016;137:696–706.
14. Bala MM, Akl EA, Sun X, Bassler D, Mertz D, Mejza F, Vandvik PO, Malaga G, Johnston BC, Dahm P, Alonso-Coello P, Diaz-Granados N, Srinathan SK, Hassouneh B, Briel M, Busse JW, You JJ, Walter SD, Altman DG, Guyatt GH. Randomized trials published in higher vs. lower impact journals differ in design, conduct, and analysis. *J Clin Epidemiol*. 2013;66:286–95.
15. Lubsen J, Kirwan B-A. Combined endpoints: can we use them? *Stat Med*. 2002;21:2959–70.
16. Pusic AL, Klassen AF, Scott AM, Klok JA, Cordeiro PG, Cano SJ. Development of a new patient-reported outcome measure for breast surgery: the BREAST-Q. *Plast Reconstr Surg*. 2009;124:345–53.
17. Carey TS, Boden SD. A critical guide to case series reports. *Spine (Phila Pa 1976)*. 2003;28:1631–4.
18. de Jong D, Vasmel WLE, de Boer JP, Verhave G, Barbé E, Casparie MK, van Leeuwen FE. Anaplastic large-cell lymphoma in women with breast implants. *JAMA*. 2008;300:2030–5.
19. Akl EA, Briel M, You JJ, Lamontagne F, Gangji A, Cukierman-Yaffe T, Alshurafa M, Sun X, Nerenberg KA, Johnston BC, Vera C, Mills EJ, Bassler D, Salazar A, Bhatnagar N, Busse JW, Khalid Z, Walker S, Cook DJ, Schunemann HJ, Altman DG, Guyatt GH. LOST to follow-up Information in Trials (LOST-IT): a protocol on the potential impact. *Trials*. 2009;10:40.
20. Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *BMJ*. 1996;312:742–4.
21. Van Allen EM, Lui VWY, Egloff AM, Goetz EM, Li H, Johnson JT, Duvvuri U, Bauman JE, Stransky N, Zeng Y, Gilbert BR, Pendleton KP, Wang L, Chiosea S, Sougnez C, Wagle N, Zhang F, Du Y, Close D, Johnston PA, McKenna A, Carter SL, Golub TR, Getz G, Mills GB, Garraway LA, Grandis JR. Genomic correlate of exceptional erlotinib response in head and neck squamous cell carcinoma. *JAMA Oncol*. 2015;1:238–44.



Quality Improvement and Patient Safety in Surgery

20

Jessica H. Hannick and Martin A. Koyle

Abbreviations

ACS	American College of Surgeons
CPGs	Clinical Practice Guidelines
CPI	Continuous Process Improvement
GRADE	Grades of Recommendation, Assessment, Development and Evaluation
HROs	Highly Reliable Organizations
IHI	Institute for Healthcare Improvement
IOM	Institute of Medicine
NSQIP	National Surgical Quality Improvement Program
PDSA	Plan, Do, Study, Act
QI	Quality Improvement
QIPS	Quality Improvement and Patient Safety
SQUIRE	Standards for Quality Reporting Excellence

Clinical Scenario

As a new pediatric urologist in a growing community hospital, Dr. H has been tasked with developing the hospital's pediatric robotic surgery program. The current staff and trainees are unfamiliar with robotic surgery and have no existing workflow for use of their expensive new robot. The first several robotic operative days are long and tedious with extensive room turnovers and frustration from all staff members. Dr. H's department chair asks why she is not performing more robotic cases each day to justify the hospital's purchase of the robot. Having just completed a fellowship, she is frustrated with the operating room (OR) inefficiencies and her inability to complete cases like they seemingly were done while in training. She recalls a seamless flow of the patient and surgical team throughout the patient's surgical journey, and wonders what can be done in her current hospital to mirror this. She wonders what other surgeons in her shoes have done faced with similar challenges and how she can learn to implement these tools at her new institution.

J. H. Hannick
Department of Urology, Cleveland Clinic Children's Hospital, Cleveland, OH, USA
e-mail: hannicj@ccf.org

M. A. Koyle (✉)
Professor Emeritus, Temerty Faculty of Medicine and Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, ON, Canada

Literature Search

The scenario above portrays a situation that is relevant to quality improvement and patient safety (QIPS). The young surgeon wishes to provide quality, safe care to the patient while also

optimizing her and her hospital's ability to do this on a daily basis for as many patients as possible. She therefore pursues a comprehensive search strategy (see [Appendix 1](#)) (See Chap. 4). PubMed guides her to a recently published quality improvement study "Increase in robot-assisted operating room efficiency: A quality improvement study" [1]. She leans in to read more about how she can use this work to tackle the problems her institution is facing.

Introduction

The practice of medicine continues to evolve, due to advances, often disruptive in nature, in basic science, pharmaceuticals and technology. However, the 'we practice is also rapidly changing, as a result of the demands and expectations of the environment(s) in which we work (government, private sector, academic, hospital), and newer approaches for whom we care. The paternalistic model of care, where a physician dictated diagnostic and management plans, has become a ' , wherein transparency and joint interaction have become the new norm. Most of us were brought up in an era of volume-based healthcare delivery. This is now the era of ' medicine, with ' care and ' becoming the pivotal goals. Value (V)-based care, will be evaluated based on performance, that is, on quality (Q) and cost (\$) ($V = Q/\$$) [2]. This approach will likely reimburse, or reward providers based on outcomes (quality) rather than on number of cases performed. With skyrocketing healthcare costs, preventable errors have become a target for QIPS [3].

For surgeons, in particular, this constantly changing playing field has to be balanced with patient safety concerns. Yet surgery has many opportunities to be successful in QIPS initiatives due to the establishment of centrally based registries, standardization, when possible, simulation training, and video critiquing of surgeries we perform. To many of us, QIPS is viewed as top-down, where institutions promote measures that we often consider nuisances and hence disregard, that are focused on issues such as hand hygiene, start times, checklists, line, surgical site and

catheter-associated urinary tract infections, falls and pressure sores, and ventilator-associated pneumonia. It, therefore, becomes incumbent upon us to define reliable (the right) and meaningful metrics to make QIPS initiatives relevant and to provide evidence that indeed we are delivering quality care.

Background

In essence, surgeons have always pursued the concept of gradual, progressive change leading to ongoing opportunities for improvement. Our surgical results are audited to search for root causes of untoward outcomes caused by errors of omission and commission related to our interventions. At the turn of the twentieth century, Ernest Amory Codman began tracking and reporting his "end results," a concept that was totally foreign to the established (Harvard) community at Massachusetts General Hospital (MGH). Codman promoted the ' , through which he would document patient demographics and track their outcomes. He suggested that it was our ethical duty to report each surgeon's results, good or bad, and identify opportunities for improvement. His desire to formally assess surgeon competence, rather than promotion based on seniority, led to his dismissal from MGH. Defiantly, he built his own facility, *The End Result Hospital*. There, he could measure efficiency and performance and present it to the public [4, 5]. To the medical community at that time, it was heresy. Ultimately, the concept of frank, scheduled discussion sessions, where analysis of hospital reform and patient outcomes occurs, has led to our now sacrosanct morbidity and mortality (M and M) conferences. He and other leaders of the newly formed American College of Surgeons (ACS) pursued a goal of evaluation and standardization. This culminated in the formation of the Joint Commission on Accreditation, for which Codman is now given much of the credit. Most recently, the ACS has focused on QIPS by developing the National Surgical Quality Improvement Program (NSQIP), which has rapidly expanded beyond North America [6]. NSQIP allows vali-

dated risk adjusted outcomes of surgical procedures to be compared between reporting hospitals. Because individual hospitals are often challenged when it comes to analyzing their surgical outcomes, it is difficult to pinpoint areas of concern to improve problem areas. NSQIP provides a tool to allow hospitals to focus on areas where improvement is necessary, where complication rates can be improved.

QIPS has now become a center of focus in all facets of medicine. It has become a mandated component of every hospital's annual report and a target for a broad variety of stakeholders, including payers and consumers (patients). The Institute of Medicine (IOM) has proposed six principles for improvement: safe, effective, patient-centered, efficient, timely, and equitable care [7]. There is even a structure for publishing, the SQUIRE (Standards for Quality Reporting Excellence) guidelines, as the increasing number of medical journals either entirely devoted, or at least partially devoted to this subject, publish new contributions [8].

What Is Quality Improvement (QI)?

Although there are many interpretations of what QI is, Batalden and Davidoff have simplistically defined it as "... combined and unceasing efforts of everyone—healthcare professionals, patients and their families, researchers, payers, planners and educators—to make the changes that will lead to better patient outcomes (health), better system performance (care) and better professional development (learning)" [9]. Berwick and colleagues at the Institute for Healthcare Improvement (IHI) developed the triple aim of quality healthcare as that of improving the experience of care, improving the health of populations, and reducing per capita costs of healthcare [10].

Many of the current concepts of continuous process improvement (CPI) are based on the work of W. Edwards Deming, his "system of profound knowledge" and his "fourteen points of management" [11]. In QI, the four keys of the system of profound knowledge, appreciating the

system, understanding variation, a theory of knowledge (epistemology), and understanding human behavior/psychology, are pivotal in instituting sustained positive changes. Simplistically, this means viewing an organizational system as a series of interrelated processes not silos where they share a common goal, understanding that variation may be either due to a common or a special cause, knowing that new knowledge is likely to be present within the organization and that there may be different individual drivers to enhance motivation. A key to Deming's thinking was that robust leaders who could engage stakeholders were mandatory, and that unwanted variation must be reduced. Kilo and IHI adopted the model for improvement in which an aim identifies a change concept to be improved within a defined time period [12]. Through a series of cycles called P-D-S-A (Plan, Do, Study, Act), a concept for change is addressed with an action plan, and an analysis of the effect of that action is then performed to allow further subsequent opportunities [13]. With each ensuing PDSA, or change cycle, small adjustments can be made with hopes of making escalating gains. PDSA reiterates three fundamental questions in each cycle: (1) What are we trying to accomplish? (2) How will we know that a change is an improvement? (3) What changes can we make that will result in an improvement? Thus, through CPI, a provider, organization, or system is more likely to improve on an ongoing basis. Other tools such as Lean/Six Sigma may be used in furthering CPI. By reducing waste, one is increasing value for potential stakeholders [14].

Is QI Research?

Both QI and traditional research are based on data. Although it may be folklore as to the originator of the statement, one of the most famous quotations attributed to Deming (and for that matter others) is "In God we trust, all others must bring data." Simplistically, research attempts to provide new knowledge while QI takes existing knowledge and seeks improvement. Not that QI lacks rigor, but primarily it is based on rapid

cycle tests of change, PDSA cycles. Although statistical methodology can be used in QI, data are best presented visually in real time using run charts or when control limits are required, Shewart (Control) charts. Because of the confusion and inconsistencies regarding the publication of QI Material, the SQUIRE guidelines were introduced in 2008.

Unique to QI reporting in healthcare are the SQUIRE guidelines [8]. SQUIRE consists of a checklist composed of 19 items/4 categories, which authors should consider when submitting papers describing new QIPs and value knowledge in healthcare. The original descriptions of SQUIRE stressed that these checklist items are common to scientific reporting but with modifications that make them more applicable to the unique aspects of medical QI work. It was acknowledged that it was important to present a system such as this as many peer-reviewed journals and stakeholders were unfamiliar with the design, implementation, and reporting of QI projects [15]. In no way are they meant to be exclusive of other guidelines and at times may be synergistic.

Discussion

Healthcare is a dynamic, complex entity that is unpredictable at best. Historically, QIPS in healthcare has been driven by external forces (National Health Services (NHS), Centers for Medicare and Medicaid Services (CMS), and Joint Commission on Accreditation of Hospital Organizations (JCAHO). As the stakes and emphasis in healthcare change and costs soar, many QIPS initiatives are now central to organizations (ACS), hospitals (Virginia Mason), and providers (Intermountain Healthcare, Kaiser-Permanente). Healthcare has attempted to emulate other highly reliable organizations (HROs) such as aviation and the nuclear agencies, where safety is prioritized and reduced errors are a priority. As value becomes more important, future research and leaders in the field will need to promote our knowledge of QIPS. In 1966, Donabedian identified three approaches in the

assessment of quality: structure, process, and outcome [16]. To this day, it remains one of the most frequently cited articles in healthcare and is a solid pillar in the evaluation of QI in healthcare [17]. In this seminal paper, structure referred not to the settings but to the qualifications of the providers in those settings and administrative systems within. Process and outcome are more straightforward and related to components of care and fate upon recovery. The Institute of Medicine further promoted QIPS in their 2000 and 2001 publications 'and' [7, 18]. The focus of these reports was that systems need to be improved knowing that human fallibility and error are innate. Thus, much harm within healthcare was preventable. With the escalation of healthcare costs, prevention of costly errors, while reducing the onus on human blame, has become a goal as the road toward HROs is traveled.

It has been estimated that approximately half of all hospital complications occur within the operating room itself [19]. Surgery is unique, as it demands the combination of one's surgical skill, knowledge, and the judgment to apply the right operation to the right patient at the right time. Addressing the fact that surgery is a team sport, exacerbates this daunting challenge. Although an operating room is as inanimate as a cockpit or an aircraft carrier, the variability in the supporting personnel (resident, fellow nursing, anesthesiologist) are not always as reliably interchangeable as those members of the airline crew or naval vessel. Although it is unlikely that the surgeon him/herself will change during an operation, commonly the other team players will. If anything, the value of standardization and checklists should not be underscored. Communication failures have been demonstrated to occur every 7–8 min during an operation [20]. Thus, the longer the operation, with increasing likelihood of greater personnel turnover, the more the opportunity for error increases. Surgical checklists are now a decade old and have been implemented to reduce 'surgically related morbidity (and mortality) [21]. Conceptually, this makes sense to assure that indeed it is the right patient who is consented for that scheduled operation, where laterality is

confirmed as are issues of allergies and antibiotic administration. Urbach, using a large Canadian provincial database, has appropriately demonstrated that there are indeed potential challenges with checklists when there is noncompliance and inconsistencies with implementation [22, 23]. Although checking boxes per se may not impact the occurrence of adverse events, improved communication among operating room personnel, regardless of “status,” is to be encouraged.

Medicine as a whole, but surgery in particular, has housed a culture of ‘where an individual, rather than a system has been held responsible for a given outcome (especially a complication). Lucien Leape, a well-known pediatric surgeon who is now at the Harvard School for Public Health, testified before the US Congress in its hearings on healthcare quality improvement that, “The single greatest impediment to error prevention in medicine is that we punish people for making mistakes.” As a result of medical error, there is often more than just the patient as a victim, as it is easier to blame an individual than a system. The IOM’s ‘*Human* [18] focuses on the fundamental issue of human fallibility. In QIPS, James Reason’s ‘has often been used as an analogy [24]. Reason uses several layers of Swiss cheese with latent and active holes in this model. In an ideal system, errors are trapped between the layers, as the holes do not align. However, given the worst-case scenario, the holes of each layer align, and this creates an opportunity for an untoward event to occur.

Recognizing that human factors play an important role in outcome, it is obvious that an understanding of other elements either support or hinder an individual’s performance. As a core element of QIPS, ‘, often called ergonomic or human engineering uses physical and psychological principles to affect processes and systems and promote a reduction in errors [25]. In the operating room setting, it has especially been applicable to medical devices, medication safety, and Information Technology.

Clinical Practice Guidelines (CPGs) have increasingly become part of our daily practices in an attempt to enhance the quality of care. To be relevant, CPGs should be based on systematic

reviews (SR) of current best evidence regarding a given topic, which has been addressed by an unbiased, multidisciplinary group of experts, so that they can be used to assist practitioners and their patients in optimizing patient care. Sackett’s description of Evidence-Based Medicine (EBM) of CPG’s, “The conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients,” is most apt [26]. Problematically, is that not all CPG’s have been based on high-quality evidence, which potentially can lead to less-than-ideal results, and thus has led to skepticism regarding their implementation. CPGs should take voluminous or complex data that is not at our fingertips, and make such data more manageable for that individual patient, who is cared for by that individual physician, at that given time. Hence, in no way should CPGs be construed as cookbook medicine or the only option for that specific patient or encounter. Sadly, defensive medicine and “over testing” are realities of the practice of medicine. This is costly to the system as a whole and impacts the utility of CPGs and other progressive initiatives such as “Choosing Wisely” [27]. In QIPS, various tools have been developed to construct and to analyze CPGs. GRADE (Grades of Recommendation, Assessment, Development, and Evaluation) has been developed as a tool for constructing CPGs that enhances physician evaluation of the quality of medical evidence so that recommendations can be applied most appropriately in patient care [28]. It replaces letter recommendations regarding quality with “strong” or “weak/conditional,” “for” or “against” and avoids “expert opinion” in the hierarchy of evidence pyramid, while recognizing that even the highest evidence standard, SR, can be flawed. Recommendations using GRADE ideally should be actionable and avoid restating obvious facts and vagaries. The AGREE II tool (Appraisal of Guidelines for Research and Evaluation) has gained acceptance as a standardized technique for the evaluation and comparison of existing CPGs [29]. It consists of 6 quality domains containing 23 items used to quantify each CPG. It requires several appraisers who independently score each domain of each guide-

line. Importantly, each appraiser also assesses the overall quality of the CPGs and makes a recommendation as to whether that guideline should be used. CPGs should be used judiciously, remembering that medical judgment and experience remain essential components of our fostering shared decision making and optimal care. CPGs will increasingly be available to us. As such, practitioners must understand how they are developed and what their qualities and their relevance are [30]. Electronic Health Records likely will improve their availability, as well as be an adjunct in improving compliance with their utilization.

Resolution of the Clinical Scenario

After reading the aforementioned article, the young surgeon incorporates numerous principles of automation and standardized work to reduce inefficiencies. A standardized OR staff task list outlines the major intraoperative steps of the most common operations along with the necessary staff and their required tasks. A list of potential supplies needed for each major step of the operations substantially reduces wait times for supply acquisition. Intraoperative observation generated a spaghetti diagram that was used to rearrange the OR setup, label the OR floor for alignment of the robot, and to eliminate movement inefficiencies of the OR team. By making these changes, the young surgeon is able to reduce setup and intraoperative times enough to increase her case volume from 1 case per working day to 2–3 cases per working day over 6 months.

Conclusion

QIPS is more than checklists and evidence-based guidelines. There are many tools that can be used in QI that are applicable to promote a culture of change and safety, eliminate silos, understand the basic problem(s) at hand, test changes, with continuous study of outcomes and performance, and to report pertinent findings to sustain change(s) to allow for opportunities to improve further [31].

The environment of the operating room is complex, and each patient is unique. As such the opportunity for error is heightened, as are opportunities to reduce, or ideally, prevent such occurrences.

Appendix 1

First navigate to the PubMed advanced search builder (<https://pubmed.ncbi.nlm.nih.gov/advanced/>). The search terms “pediatric urology” AND “quality improvement,” AND “robot” are then entered. The search is narrowed to English language papers, Age “Child: birth-18 years” from October 2019–October 2024. The top result using this search strategy identifies the article that addresses the question posed.

References

1. Kanabolo DL, Merguerian P, Ahn J, Low DK, Fernandez N. Increase in robot assisted operating room efficiency: a quality improvement study. *J Pediatr Urol.* 2023;S1477-5131(23):00519–3. <https://doi.org/10.1016/j.jpuro.2023.11.019>.
2. Institute of Medicine, Committee on Redesigning Health Insurance Performance Measures, Payment, and Performance Improvement Programs. *Rewarding provider performance: aligning incentives in medicare.* Washington, DC: National Academies Press; 2016.
3. Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, Etchells E, Ghali WA, Hébert P, Majumdar SR, O’Beirne M, Palacios-Derflinger L, Reid RJ, Sheps S, Tamblyn R. The Canadian adverse events study: the incidence of adverse events among hospital patients in Canada. *CMAJ.* 2004;170(11):1678–86. <https://doi.org/10.1503/cmaj.1040498>.
4. Berwick DM. EA Codman and the rhetoric of battle: a commentary. *Milbank Q.* 1989;67(2):262–7.
5. Donabedian A. The end results of health care: Ernest Codman’s contribution to quality assessment and beyond. *Milbank Q.* 1989;67(2):233–56.
6. Birkmeyer JD, Shahian DM, Dimick JB, Finlayson SR, Flum DR, Ko CY, Hall BL. Blueprint for a new American College of Surgeons: National Surgical Quality Improvement Program. *J Am Coll Surg.* 2008;207(5):777–82.
7. Institute of Medicine, Committee on Quality of Health Care in America. *Crossing the quality chasm: a new health system for the 21st century.* Washington, DC: National Academies Press; 2001.

8. Davidoff F, Batalden PB, Stevens D, Ogrinc G, Mooney S, SQUIRE Development Group. Publication guidelines for quality improvement in health care: evolution of the SQUIRE project. *Qual Saf Health Care*. 2008;17(Suppl 1):3–9.
9. Batalden PB, Davidoff F. What is “quality improvement” and how can it transform healthcare? *Qual Saf Health Care*. 2007;16(1):2–3. <https://doi.org/10.1136/qshc.2006.022046>.
10. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. *Health Aff*. 2008;27(3):759–69.
11. Walton M. The Deming management method. New York: Perigee Books; 1986.
12. Kilo CM. A framework for collaborative improvement: lessons from the Institute of Healthcare Improvement’s breakthrough series. *Qual Manag Health Care*. 1998;6(4):1–13.
13. Papanikolaou F, Goldsmith CH. Testing quality improvement interventions. *Clin Plastic Surg*. 2008;35:297–303.
14. Cima RR, Brown MJ, Hebl JR, Moore R, Rogers JC, Kollengode A, Amstutz GJ, Weisbrod CA, Narr BJ, Deschamps C. Surgical process improvement team. Use of lean and six sigma methodology to improve operating room efficiency in a high volume tertiary-care academic medical center. *J Am Coll Surg*. 2011;213(1):83–92.
15. Davidoff F, Batalden PB. Toward stronger evidence on quality improvement. Draft publication guidelines: the beginning of a consensus project. Transformation of healthcare—quality improvement. *Qual Saf Health Care*. 2005;14(5):319–25. <https://doi.org/10.1136/qshc.2005.014787>.
16. Donabedian A. Evaluating quality of medical care. *Milbank Q*. 1966;44(3, Part 2):166–206. <https://doi.org/10.2307/3348969>.
17. Ayanian JZ, Markel H. Donabedian’s lasting framework for health care quality. *New Eng J Med*. 2016;375(3):205–7. <https://doi.org/10.1056/NEJMp1605101>.
18. Kohn LT, Corrigan J, Donaldson MS. To err is human: building a safer health system. Washington, DC: National Academies Press; 2000.
19. Gawande AA, Thomas EJ, Zinner MJ, Brennan TA. The incidence and nature of surgical adverse events in Colorado and Utah in 1992. *Surgery*. 1999;126(1):66–75.
20. Hu YY, Arriaga AF, Peyre SE, Corso KA, Roth EM, Greenberg CC. Deconstructing intraoperative communication failures. *J Surg Res*. 2012;177(1):37–42.
21. Haynes AB, Weiser TG, Berry WR, Lipsitz SR, Breizat AH, Dellinger EP, Herbosa T, Joseph S, Kibatala PL, Lapitan MCM, Merry AF, Moorthy K, Reznick RK, Taylor B, Gawande; Safe Surgery Saves Lives Study Group. A surgical safety checklist to reduce morbidity and mortality in a global population. *N Engl J Med*. 2009;360(5):491–9.
22. Urbach DR, Govindarajan A, Saskin R, Wilton AS, Baxter NN. Introduction of surgical safety checklists in Ontario, Canada. *N Engl J Med*. 2014;370(11):1029–38.
23. Leape LL. The checklist conundrum. *N Engl J Med*. 2014;370(11):1063–4.
24. Reason J. Human error models and management. *Br Med J*. 2000;320:768–70.
25. Vosper H, Hignett S, Bowie P. Twelve steps for embedding human factors and ergonomics principles in healthcare education. *Med Teach*. 2018;40(4):357–63.
26. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn’t. *BMJ*. 1996;312(7023):71–2. <https://doi.org/10.1136/bmj.312.7023.71>.
27. Welk B, Winick-Ng J, McClure JA, Lorenzo AJ, Kulkarni G, Ordon M. The impact of the choosing wisely campaign in urology. *Urology*. 2018;116:81–6. <https://doi.org/10.1016/j.urology.2018.03.005>.
28. Guyatt GH, Oxman AD, Vist G, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924–6. <https://doi.org/10.1136/bmj.39489.470347.AD>.
29. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Fedar G, Fervers B, Graham ID, Grimshaw J, Hanna SE, Littlejohns P, Makarski J, Zitzelsberger L, for the AGREE Next Steps Consortium. AGREE II: advancing guideline development, reporting, and evaluation in health care. *CMAJ*. 2010;182(18):E839–42. <https://doi.org/10.1503/cmaj.090449>.
30. Koyle MA. Truth and consequences—the issue with standards and guidelines. *Can Urol Assoc J*. 2018;12(4):79–80.
31. Koyle MA, Koyle LCC, Baker GR. Quality improvement and patient safety: reality and responsibility from Codman to today. *J Pediatr Urol*. 2018;14(1):16–9.



Diagnostic Studies in Surgery

21

Michael Xie, Stuart Archibald, Achilles Thoma,
Michael K. Gupta, Jessica Murphy,
and Charles H. Goldsmith

Abbreviations

ACR	American College of Radiology
FN	False Negative
FNAB	Fine Needle Aspiration Biopsy
FP	False Positive
LR	Likelihood Ratio
NPV	Negative Predictive Value
PPV	Positive Predictive Value
TIRADS	Thyroid Imaging Reporting and Data System
TN	True Negative
TP	True Positive
US	Ultrasound

Introduction

Surgeons are challenged every day in clinical practice to make a diagnosis. A correct diagnosis is foundational to recommending the proper management, whether that is to offer surgery or to offer serial observation and refrain from surgery. For example, a common clinical dilemma faced by surgeons is the management of thyroid nodules. Thyroid nodules are portions of the thyroid gland that appear distinct from the normal thyroid parenchyma on imaging. The standard investigation of thyroid nodules includes ultrasound (US) of the thyroid and fine-needle aspiration biopsy (FNAB). Intermediate or indeterminate results are common in the investigations. A comprehensive approach to these indeterminate results requires that multiple management options be explored in a shared decision-making process.

M. Xie · S. Archibald · M. K. Gupta (✉)
Division of Head and Neck Surgery, Department of Surgery, McMaster University,
Hamilton, ON, Canada
e-mail: michael.xie@medportal.ca;
archibs@mcmaster.ca; mgupta@mcmaster.ca

A. Thoma
Division of Plastic Surgery, Department of Surgery & Department of Health Research Methods, Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: athoma@mcmaster.ca

J. Murphy
Division of Physical Medicine & Rehabilitation, Department of Medicine, McMaster University,
Hamilton, ON, Canada
e-mail: murphj11@mcmaster.ca

C. H. Goldsmith
Department of Health Research Methods, Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: goldsmi@mcmaster.ca

The present climate, in which patients often come primed with internet-derived “expertise,” can add social complexity and time to the consultation process. Furthermore, patients have continually rising expectations for certainty in the diagnosis before making their decision.

In this chapter, using a clinical scenario of an informed patient who has typical expectations, we will demonstrate how to use the best evidence to find, evaluate, and apply diagnostic tests.

Clinical Scenario

A 42-year-old woman is referred to you with an asymptomatic thyroid nodule found on US, which was performed to investigate vague swallowing symptoms. She had no history of previous thyroid disease or head and neck radiation exposure. However, she is alarmed with the identification of the nodule because her grandmother died years ago of a thyroid cancer. She is “surgery averse,” and only wants an operation “if it is necessary.” Accordingly, she performed an Internet search and discovered a molecular test, ThyroSeq ver3, which, it is claimed, will better determine if her lesion is malignant when performed on a FNAB sample.

Before looking at her printed handout, you examine her; the examination discloses only that there is a fullness to her right lower neck in the thyroid compartment, without any well-defined mass. No neck adenopathy could be palpated. Laryngoscopy showed that her vocal cords moved normally. The remainder of the examination is noncontributory.

This negative history and non-palpable nodule cannot influence the probability of this patient having cancer, thus her probability of malignancy remains at the baseline prevalence of thyroid cancer in the general population. You know that the prevalence of thyroid nodules and thyroid cancer has increased in recent years, with most of this increase considered to be due to the increased and widespread use of US examinations [1, 2]. This constitutes “detection bias,” where the event of interest (thyroid nodules and thyroid cancer) is observed more frequently in the group that has a

more sensitive detection method, in this case, US, applied to it than is applied to the other group; this may lead to the mistaken interpretation that nodules and cancers are truly increasing [3]. You believe that this patient falls into this category.

Nodules may be found in 70% of the US population [4], and about 5% of thyroid nodules are malignant [5, 6], thus a reasonable estimate for the prevalence of thyroid cancer is 3%. On review of this patient’s US report, unfortunately, there is very little additional information on the characteristics of the nodule aside from its dimensions. In your own institution, US reports are given in the TIRADS (Thyroid Imaging Reporting and Data System) format, which provides more information than was given in this patient’s study. You decide to search the literature to find out what are the current recommendations for diagnostic tests on thyroid nodules, their sequence, and the role of TIRADS.

Literature Search

As outlined in Chap. 3, the first step to solving a clinical dilemma is to form your research question using the PICOT format (Box 21.1).

Box 21.1 The PICOT Format Applied to the Current Clinical Scenario

Population	Patient with <i>thyroid nodules</i>
Intervention	TIRADS (Thyroid Imaging Reporting and Data System)
Comparative intervention	NA
Outcome	Risk classification for cancer diagnosis
Time horizon	NA

Using the following search terms, you perform a literature search using Cochrane Library: “Thyroid Nodules” AND “TIRADS”. Your search yields 18 articles, of these 2 studies and 1 active trial examine the diagnostic performance of a thyroid US risk stratification system (see Appendix 1); however, none of these articles are

ideal as the studies examine alternative regional variations of TIRADS, and therefore may not be generalizable to your patient and center. The American College of Radiology (ACR) TIRADS was developed in North America and implemented at your center, therefore, it may be more generalizable and applicable to this clinical scenario. You decide to perform a literature search using the terms “Thyroid Nodules” AND “TIRADS” AND (“American College of Radiology” OR “ACR”) in OVID Medline, and this time your search yields 99 articles. By restricting your search to publications within the last 3 years (2021–2024), your search results decrease to 63 ([Appendix 2](#)). Out of these studies, you choose a 2021 paper by Huh et al. [7]; you choose this study as it reported data on American College of Radiology (ACR) TIRADS in a large sample of nodules and utilized a pragmatic approach for defining benign versus malignant reference.

Brief Study Summary

It is clear that the most commonly recommended initial test of a thyroid nodule is US of the thyroid. Multiple methods of thyroid US interpretation exist including the American Thyroid Association Guideline Sonographic Patterns [8] as well as regional variations of TIRADS [9]. The use of TIRADS categories has become dominant, in particular the ACR-TIRADS in North America and your institution.

In this study, the investigators retrospectively reviewed 2106 thyroid nodules 10 mm or larger in size with definitive cytopathologic diagnosis and compared it to US ACR- and K-TIRADS

results as the index test. The ACR-TIRADS threshold for clinical significance (i.e., warrants fine needle aspiration biopsy [FNAB]) was further stratified by size criteria within each TIRADS category per ACR-TIRADS (≥ 2.5 cm for TIRADS 3, ≥ 1.5 cm for TIRADS 4, ≥ 1.0 cm for TIRADS 5) [10]. The cytopathologic reference standard was defined as benign or malignant based on definitive cytologic (Bethesda II, VI) or histopathologic (benign or malignant tissue pathology) diagnosis. Indeterminate cytologic diagnoses including nondiagnostic, atypia or follicular lesion of undetermined significance, follicular neoplasm, or suspicious for malignancy without further cytologic or pathologic diagnostic clarification were excluded in this study. Other studies have utilized surgical pathology only as the reference standard, which results in selection bias against benign/nonsurgical thyroid nodules, or FNAB alone as the reference standard, which includes an “indeterminate” category where only some nodules come to surgery and the actual status of some nodules remain uncertain, thus lacks a consistent reference standard.

In the current study by Huh et al. [7], all patients between March 2017 and January 2019 undergoing US and FNAB for thyroid nodules were reviewed. The study examined 3149 thyroid nodules and 1043 nodules were excluded for indeterminate cytopathologic diagnosis. US and prospective US characteristic reporting at the time of biopsy were performed by one of twelve board certified radiologists. A single, blinded, experienced radiologist independently assigned an ACR- and K- TIRADS score/category to each nodule. In total, 2084 patients, and 2106 nodules were studied. [Figure 21.1](#) outlines the study design by Huh et al. [7].

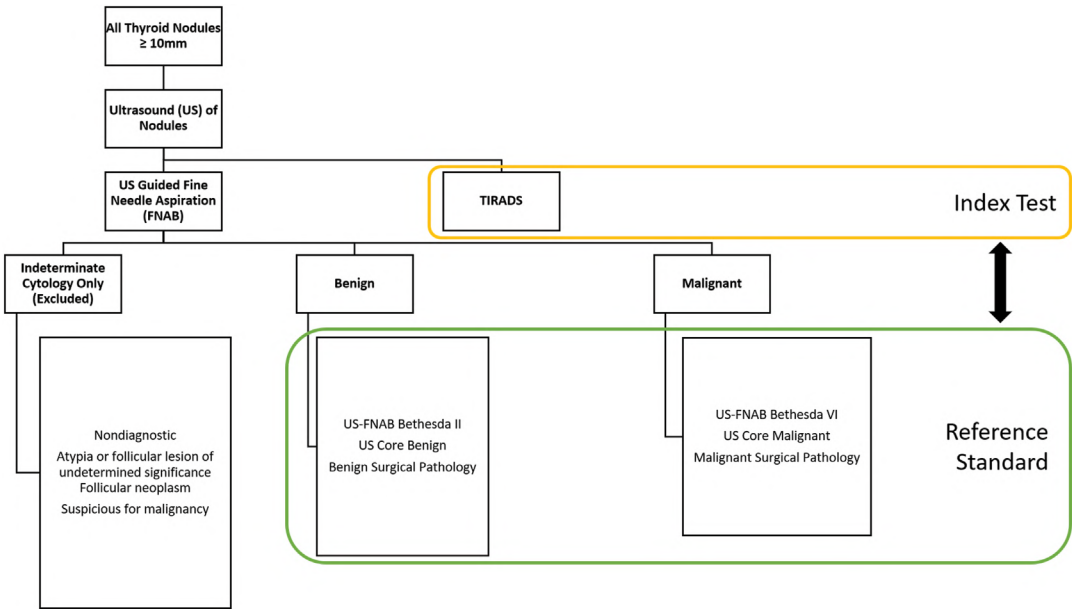


Fig. 21.1 Flow chart of study design. (Created using information from Huh et al. [7])

Important Aspects of Diagnostic Studies

The first step in successfully understanding and appraising a diagnostic study is to understand the terms commonly used in this area of research. These terms are summarized in Box 21.2. As also seen in Chap. 22, diagnostic studies use the terms “True Positive” (TP), “False Positive” (FP), “True Negative” (TN), and “False Negative” (FN). True Positive, refers to when the test correctly diagnoses the presence of a disease, whereas False Positive, incorrectly diagnoses an individual with a disease when they do not have the disease. Similarly, a True Negative is when a test correctly determines that a disease is not present, and a False Negative refers to when a diagnostic test determines a patient does not have a disease, when in fact they do. These terms, which describe how useful test is, are outlined in terms of cancer diagnosis in Table 21.1.

When a clinician is making a diagnosis (Box 21.3), the direction of reasoning moves from a test result to the probability of disease.

The Likelihood Ratio (LR), a computation not commonly used by surgeons (although it should be), does this. It can be either “LR+” or “LR–” depending on whether or not one is considering a positive test or a negative test with exactly two levels. An LR+ test answers the question: “What is the probability that this positive test is found in a patient with the disease compared to the chance that this same positive test is found in a patient without the disease?” Referring to Table 21.1, where the data are displayed in a 2 × 2 table,

Table 21.1 Diagnostic testing and true results in cancer diagnosis

	Cancer Present	Cancer Absent
Test Positive for Cancer	True Positives (TP) ^a a ^b	False Positives (FP) b
Test Negative for Cancer	False Negatives (FN) c	True Negatives d
Totals	(TP + FN) (a + c)	(FP + TN) (b + d)

^a“TP,” “FP,” “FN,” and “TN” are used to designate the *cell*
^b“a,” “b,” “c,” and “d” are used to designate the *numerical values* in a cell

Box 21.2 Important Terms for Understanding Diagnostic Tests

- *True Positive (a)*
The test indicates a positive result AND the reference standard indicates a positive result
- *True Negative (d)*
The test indicates a negative result AND the reference standard indicates a negative result
- *False Positive (b)*
The test indicates a positive result BUT the reference standard indicates a negative result
- *False Negative (c)*
The test indicates a negative result BUT the reference standard indicates a positive result
- *Positive Predictive Value (PPV): $a/(a + b)$*
The proportion of positive tests that is correct
- *Negative Predictive Value (NPV): $d/(c + d)$*
The proportion of negative tests that is correct
- *Sensitivity: $a/(a + c)$*
The proportion of patients with disease that is correctly identified by the test Synonyms: True-Positive Rate or Positivity in Disease
- *Specificity: $d/(b + d)$*
The proportion of patients without the disease correctly identified by the test Synonyms: True-Negative Rate or Negativity in Health
- *Likelihood Ratio (LR):*
Ratio of the probability that a test result is correct to the probability that the test result is not correct
- *Likelihood Ratio of a Positive Test (LR+)*
The equivalent of dividing the true-positive rate by the false positive rate, or $\frac{\text{Sensitivity}}{1 - \text{Specificity}}$ ^a
- *Likelihood Ratio of a Negative Test (LR-)*
The equivalent of dividing the false negative rate by the true-negative rate, or $\frac{1 - \text{Sensitivity}}{\text{Specificity}}$ ^a
- *“Accuracy” (best avoided) = $(a + d)/(a + b + c + d)$*

	Reference standard/disease is positive/present	Reference standard/disease is negative/absent	Total
Test is positive	<i>a</i>	<i>b</i>	<i>a + b</i>
Test is negative	<i>c</i>	<i>d</i>	<i>c + d</i>
	<i>a + c</i>	<i>b + d</i>	<i>a + b + c + d</i>

^aFor two-level tests

$$\text{LR+ is calculated as : LR+} = \frac{\text{Probability of a positive test in those with cancer}}{\text{Probability of a positive test in those without cancer}} = \frac{\text{sensitivity}}{(1 - \text{specificity})}$$

LR− answers the question: “What is the probability that this negative test will be found in a patient with the disease compared to the proba-

bility that this negative test will be found in a patient without the disease?”

$$\text{LR− is calculated as :} = \frac{\text{Probability of a negative test in those with cancer}}{\text{Probability of a negative test in those without cancer}} = \frac{(1 - \text{sensitivity})}{\text{specificity}}$$

The above is applied to the results from Huh et al. [7], as follows:

$$\text{LR+} = \frac{a/(a+c)}{b/(b+d)} = \frac{451/522}{584/1584} = 2.34$$

This is how Huh et al. [7] used the data, which meant reducing it to two levels based on ACR-TIRADS category and size criteria to get a “pooled LR”.

$$\text{LR−} = \frac{c/(a+c)}{d/(b+d)} = \frac{71/522}{1000/1584} = 0.22$$

LR also can be applied to a single level out of a multilevel data set, avoiding the need to collapse the data into a 2×2 table, which results in a loss of information. The data from Huh et al. [7] is shown for each level in Table 21.2. For simplicity, size criterion was not accounted for in the LR calculation.

When an LR is 1.0, the pretest probability (prevalence), and posttest probability are unchanged. When an LR is greater than 1.0, then the posttest probability of the diagnosis has been *increased* from the pretest probability (prevalence), and conversely when an LR is less than 1.0, then the post-chance probability of the diagnosis has been *decreased* from the pretest probability (prevalence). Generally, LRs greater than 10 indicate large and likely conclusive changes in the pretest probabilities and “rule in” the condition of interest, whereas LRs less than 0.1 indicate large and likely conclusive changes in the pretest probabilities and “rule out” the condition of interest. LRs between 2 and 5, or between 0.5 and 0.2 indicate modest but possibly important changes in the pretest probabilities. LRs between 1 and 2, or between 0.5 and 1.0 indicate small and not usually conclusive changes in pretest probabilities [11]. The additional advantages of LRs are that the value is not influenced by preva-

Table 21.2 Results as reported in the supplementary data by Huh et al. [7] showing LR for each category

Test result (ACR-TIRADS)	Malignant	Benign	Totals	LR+
1	4	162	166	0.075
2	8	427	435	0.057
3	14	463	477	0.092
4	93	422	515	0.67
5	403	110	513	11.11
Total	522	1584		

lence of the condition. LRs can be used sequentially in several tests, which follows the pattern of use in clinical practice. Thus, the pretest probability and LR can give the posttest probability of the first test, which serves as the pretest probability with the LR to give the posttest probability on the second test, and so on. This can be done using mathematical formulae for each test, or more simply by applying a nomogram to visually give the results [12]. This nomogram, shown in Fig. 21.2, makes it easy to use cascading tests in this manner [11]. By placing a ruler on the pretest probability and aligning it with the calculated LR we can obtain on the far right of the nomogram the posttest probability of the disease. As TIRADS in general has multiple categories and different size thresholds, various cut-off points may be selected for determining sensitivity, specificity, positive predictive value, negative predictive value, and likelihood ratios. The cut-off point shown in Table 21.3 in the Huh et al. [7] article is at ≥ 2.5 cm for TIRADS 3, ≥ 1.5 cm for TIRADS 4, and ≥ 1.0 cm for TIRADS 5; thus TIRADS 1, 2, and 3–5 not meeting size criteria, are considered “negative” for cancer, and TIRADS 3–5 meeting size criteria are considered “positive” for cancer.

Sensitivity and specificity are commonly the stated measures of a diagnostic test, even though they reverse the process used clinically by using

the known disease categories to get test performance measures. The term “sensitivity” describes the proportion of diseased individuals in the population that are classified as having the disease

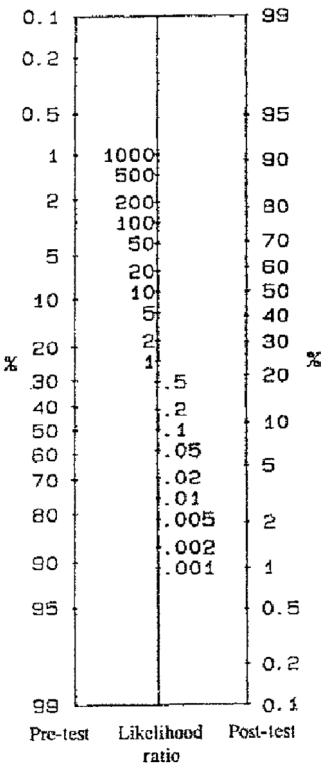


Fig. 21.2 Nomogram for interpreting diagnostic test result

Table 21.3 Results as reported by Huh et al. [7] grouping TIRADS categories. Pooled LR+ = 2.34

	Malignant (US-FNAB Bethesda VI, US Core Malignant, Malignant Surgical Pathology)	Benign (US-FNAB Bethesda II, US Core Benign, Benign Surgical Pathology)	Totals
“Positive” ≥ 2.5 cm, TIRADS $3 \geq 1.5$ cm, TIRADS $4 \geq 1.0$ cm, TIRADS 5	451 (TP) <i>a^b</i>	584 (FP) <i>b</i>	1035 <i>a + b</i>
“Negative” TIRADS 1 TIRADS 2 <2.5 cm, TIRADS $3 < 1.5$ cm, TIRADS $4 < 1.0$ cm, TIRADS 5	71 (FN) <i>c</i>	1000 (TN) <i>d</i>	1071 <i>c + d</i>
Totals	522 <i>a + c</i>	1584 <i>b + d</i>	2106 <i>a + b + c + d</i>

^aTP True Positive; FP False Positive, FN False Negative; TN True Negative, denote the “cell names”

^b“*a*,” “*b*,” “*c*,” “*d*” refer to the numerical values of those cells: 451, 584, 71, 1000, respectively

using the test. In other words, sensitivity describes the probability that the test is able to correctly diagnose a disease. Sensitivity is calculated using the formula below (for 2×2 tables like Table 21.1):

$$\text{Sensitivity} = \frac{a}{a + c}$$

Specificity is the term used to describe the proportion of those people without the disease who were correctly identified by a negative test as non-diseased. Specificity is calculated using the formula below:

$$\text{Specificity} = \frac{d}{b + d}$$

Using results from the Huh et al. [7] study (Table 21.3), sensitivity and specificity would be calculated as:

$$\text{Sensitivity} = \frac{451}{451 + 71} = \frac{451}{522} = 0.864 \text{ or } 86.4\%$$

and

$$\begin{aligned} \text{Specificity} &= \frac{1000}{584 + 1000} = \frac{1000}{1584} \\ &= 0.631 \text{ or } 63.1\% \end{aligned}$$

While sensitivity and specificity are useful in characterizing diagnostic tests and are commonly reported, they have major limitations when applied to “real-life” practice [13]. For instance, as noted above, in using sensitivity and specificity, the clinician’s reasoning must move from the disease to the test result. But the presence or absence of the disease is precisely what is being sought in a diagnostic quest, where the clinical reasoning moves from the test result to whether the disease is present or not. Furthermore, sensitivity and specificity do not allow sequential use in tests, and thus are limited in refining the diagnosis. Sensitivity and specificity are developed from binary data, and are not amenable to being used on a single level in a multilevel data set, without reducing the data to a 2×2 table, thereby losing information. Finally, as explained by Parikh et al. [13], in the early stages of a disease, it is difficult to determine the presence of health or illness and so the sensitiv-

ity may decrease; on the other hand, in cases of severe disease, the sensitivity may increase. Despite this, for most diagnostic tests in common situations, the sensitivity and specificity are considered to be stable. For all of these reasons, the most useful test for a clinician is the LR. The reported Positive Predictive Value (PPV) and Negative Predictive Value (NPV) for 2×2 tables also answer a relevant question when making a diagnosis, but they are computed using prevalence of disease in the data set, which may not be the prevalence of disease in your clinical setting [13]. In the study by Huh et al. [7], the prevalence is $(a + c)/(a + b + c + d) = 522/2106 = 0.248$ or 24.8%. This makes the PPV and NPV really only useful at around 50% prevalence. You need to ask, “Is this true for my clinical setting?”.

You decide to review the article by Huh et al. [7] and a subsequent systematic review article by Kwak et al. [14] to see how the TIRADS system was developed and how it has evolved. You discover that the TIRADS system, as first proposed by Horvath et al. [15] in 2009, used diagnostic ultrasonography to stratify the risk of cancer into six levels. The original proposal by Horvath et al. [15] had ten US patterns to define the levels of cancer risk, and it has been validated by using surgical pathology as the “gold standard.” However, while the concept of a model stratifying cancer risk remained attractive, this TIRADS system has proven too cumbersome for most investigators, and some nodules do not fit the patterns [14]. Thus, modifications to simplify the system have been made. The first suggestion was to de-emphasize the criteria designed to characterize specific benign conditions because with most thyroid nodules, the main goal is to identify those nodules that are malignant. The second was to reduce the number of US characteristics of the nodule to the five considered relevant for determining risk of cancer: composition (solid, cystic, mixed), echogenicity (hyper, hypo, marked hypo), adverse margins (yes/no), adverse calcifications (yes/no), and adverse shape (yes/no). These suggested modifications were used to define levels 4A, 4B, 4C, and 5, where 4A has one of the US characteristics present, 4B has two,

Table 21.4 Risk of malignancy associated with each category

TIRADS	Risk of cancer [% (95% Confidence Interval)]
5	35 (31.0–39.1)
4	9.1 (7.6–10.8)
3	4.8 (3.4–6.5)
2	1.5 (0.6–2.9)
1	0.3 (0–1.8)

Created using information from Middleton et al. [16]

and 4C has three or four [14]. ACR-TIRADS further simplifies the categories down to summation of points correlating to five categories: TIRADS 1 (0 points), 2 (2 points), 3 (3 points), 4 (4–6 points), 5 (7 or more points) [13]. The risk of malignancy associated with each category is shown in Table 21.4. The effect of these modifications has been to simplify the TIRADS system and so making it easier to apply. An additional result is that “TIRADS” may mean slightly different things depending on which TIRADS model is being used.

For instance, using the modified system by Kwak et al. [14] or that of Horvath et al. [15] will have different diagnostic performances compared to ACR-TIRADS. The use of different criteria for the levels is a weakness of the present status of TIRADS. In any TIRADS system, the expertise of the US operator is important. The US operators in published studies are usually the most available experts, and this level of expertise may affect the generalizability of the results.

Once an US, using the TIRADS categories, has been found to warrant FNAB (depending on TIRADS category and size), then the patient is usually sent for an US-guided FNAB, and the results of this test are reported using the commonly accepted Bethesda System for Reporting Thyroid Cytopathology [17]. The Bethesda System uses “diagnostic categories” designated by Roman numerals, rather than Arabic numerals as with TIRADS levels. A meta-analysis of the Bethesda System by Bongiovanni et al. [18] presents data that allow LR_s to be calculated for the

Table 21.5 Bethesda diagnostic categories showing percentage of cancer. Pooled LR₊ = 1.97

Bethesda Category	Risk of cancer
I	16.8%
II	3.7%
III	15.9%
IV	26.1%
V	75.2%
VI	98.6%

various Diagnostic Categories. Thus, the diagnostic tests of US with TIRADS categories can be followed by FNAB with Bethesda Diagnostic categories, and by using LR_s in each test in the sequence, the probabilities of cancer may be refined. The LR_s for the Bethesda Diagnostic Categories calculated from data in the Bongiovanni et al. [18] study are displayed in Table 21.5.

	Malignant	Benign	Totals
“Positive” Bethesda IV–VI	1851 (TP) ^a <i>a</i> ^b	1461 (FP) <i>b</i>	3312 <i>a</i> + <i>b</i>
“Negative” Bethesda II	58 (FN) <i>c</i>	1505 (TN) <i>d</i>	1563 <i>c</i> + <i>d</i>
Totals	1909 <i>a</i> + <i>c</i>	2966 <i>b</i> + <i>d</i>	4875 <i>a</i> + <i>b</i> + <i>c</i> + <i>d</i>

Created using information from Bongiovanni et al. [18]
Bethesda categories I and III are indeterminate results that warrant further investigations for risk stratification (repeat FNA, molecular testing, diagnostic surgery). Although indeterminate, Bethesda IV is followed by diagnostic surgery in traditional thyroid nodule management [18]
^a*TP* True Positive; *FP* False Positive, *FN* False Negative; *TN* True Negative, denote the “cell names”
^b“*a*,” “*b*,” “*c*,” “*d*” refer to the numerical values of those cells: 1851, 1461, 58, 1505, respectively

Evaluating the Literature on a Diagnostic Test

Three key steps must be taken to assess a diagnostic test, and this can be done by answering the three main questions outlined in Box 21.3.

Box 21.3 Questions to Appraise the Literature on a Diagnostic Test

1. Are the results valid?

Primary guides

- (a) Is there an independent, blind comparison with a reference standard?
- (b) Does the patient sample include an appropriate spectrum of patients whom the diagnostic test will be applied in clinical practice?

Secondary guides

- (a) Do the results of the test being evaluated influence the decision to perform the reference standard?
 - (b) Are the methods for performing the test described in sufficient detail to permit replication?
2. What are the results?
- (a) Are likelihood ratios of the test being evaluated? Or, are the data necessary for their calculation provided?
3. Will the results help me in caring for my patients?
- (a) Will the reproducibility of the test result and its interpretation be satisfactory in my setting?
 - (b) Are the results applicable to my patient?
 - (c) Will the results change my management?
 - (d) Will patients be better off as a result of the test?

Created using information from Jaeschke et al. [19]

Are the Results of the Study Valid?

The answer to this question is critical because it determines whether the results can be trusted. You need to know if the test was compared to an appropriate reference standard, and what the characteristics of the study population are.

Primary guide #1(a): Is there an independent blind comparison with a reference standard?

In the study by Huh et al. [7], all eligible patients' thyroids underwent US with sonographic descriptive characterization and those ≥ 10 mm were biopsied. Indeterminate cytology without further clarification through repeat FNAB or surgery were excluded (Bethesda I, III, IV). The reference standard was defined as:

- Benign lesions were defined as Bethesda II cytology or benign tissue pathology
- Malignant lesions were defined as Bethesda VI or malignant tissue pathology.

ACR- and K- TIRADS scoring was performed post hoc by an experienced radiologist who was blinded to clinical and cytopathologic data.

Primary guide 1(b): Does the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice?

Given that the center reporting is a tertiary referral center [7], the spectrum of disease might at first be thought to be of advanced disease. However, the size of the nodules indicates that many small and therefore early nodules are included, so the spectrum seems to be applicable more broadly. Moreover, the setting is generalizable to your practice if you are working in a tertiary care setting.

Huh et al.'s [7] study sample had a broader inclusion criteria due to the size indication for biopsy at ≥ 10 mm compared to the ACR-TIRADS (≥ 2.5 cm for TIRADS 3, ≥ 1.5 cm for TIRADS 4, ≥ 1.0 cm for TIRADS 5). 10 mm is generally regarded as the threshold for possible invasive intervention as lesions smaller than this represent benign or microcarcinomas, which would be managed conservatively or with serial sonographic follow-up; therefore, this was an appropriate minimum nodule size. Data on FNAB cytology on lesions that would not otherwise have been biopsied based on ACR-TIRADS size criteria allows for determination of false positive and true negatives.

Indeterminate cytologic diagnoses including nondiagnostic, atypia or follicular lesion of unde-

terminated significance, follicular neoplasm, or suspicious for malignancy without further cytologic or pathologic diagnostic clarification were excluded in this study. Indeterminate nodules that do not have definitive cytology on repeat FNAB or subsequent surgical pathology have uncertain diagnoses and thus lack a consistent reference standard. Moreover, this subgroup of indeterminate nodules represents a separate clinical entity whereby management options for diagnostic clarification will alter the risk of malignancy and guide further management decisions.

Properties of a test can change with different degrees of disease severity. Likelihood ratios tend to increase when patients with the target disease all have severe disease and tend to decrease toward 1 when the subjects all have mild disease. Knowing the study population will help you decide if the results are applicable to your practice.

Given that the data came from a single tertiary center known for its interest in thyroid cancers, the prevalence of malignant disease in the population from which the study sample was drawn will be high. This likely accounts for the high (24.8%) percentage of remaining nodules that are malignant. The most common malignant pathologies were papillary thyroid carcinoma and minimally invasive follicular thyroid carcinoma. The most common benign pathologies were adenomatous hyperplasia and follicular adenoma. Both of these reflect the expected epidemiology of malignant and benign thyroid lesions. Staging data was not available. However, given the size of the nodules (mean 23.3 mm) and age of patients (mean 50.4), it is likely that most of the disease is early stage, with some more advanced lesions, as the largest was 10 cm. It seems therefore that an appropriate sample of cancer patients was studied.

Once validity has been confirmed, and the subjects characterized, two more questions need to be answered.

Secondary guide #1(a): Do the results of the test being evaluated (in this case TIRADS) influence the decision to perform the reference standard test(s) (in this case, FNAB and surgery)?

In this study, the TIRADS category was determined post hoc. All thyroid nodules ≥ 10 mm undergoing FNAB during the study period were included. Although it is not clearly stated in the text whether initial US (or any risk stratification) guided the decision for biopsy, there is a large sample of TIRADS 1 and 2 nodules that underwent FNAB suggesting a less restrictive indication or lower size threshold for FNAB.

Secondary guide #1(b): Are methods for performing the test described in sufficient detail to permit replication?

The performance of all of the tests (US, TIRADS scoring/categorizing, FNAB, pathology) are well-described. Examples of each of the US characteristics, reference to specific ACR- and K-TIRADS criteria, the equipment (US machine) for TIRADS, and reference to the Bethesda Classification were specified [7].

What Are the Results?

Given the above considerations, you are becoming more confident that the study's results will be believable. We have presented LR_s as the most versatile and applicable way to use tests to refine the diagnosis. LR_s answer the relevant question, use the same pattern of reasoning as does a diagnostician (unlike sensitivity and specificity), are not affected by prevalence, (unlike PPV and NPV), can be applied to multilevel data without loss of information (unlike sensitivity and specificity), and can be used on sequential tests to narrow the diagnostic probabilities.

Are likelihood ratios for the test results presented, or are the data necessary for their calculation available?

Sufficient data are presented to allow LR_s to be calculated for each TIRADS level. Referring to the previous calculations, and the Huh et al. [7] paper, the data give a pooled LR₊ of 2.34 (Table 21.3). However, pooled LR₊ of 2.34 was derived from TIRADS levels and size criteria for biopsy (≥ 2.5 cm for TIRADS 3, ≥ 1.5 cm for TIRADS 4, ≥ 1.0 cm for TIRADS 5) and therefore not simply derived as in Table 21.2. Reference standards for the study were a combi-

Table 21.6 Calculated LR_s for each category of TIRADS 4 and 5 when TIRADS 2 and 3 are discarded (Option 3)

Test result (ACR-TIRADS)	Malignant	Benign	Totals	LR+
3	14	463	477	0.059
4	93	422	515	0.43
5	403	110	513	7.15
Total	510	995		

Created with data from Huh et al. [7]

nation of definitive cytologic and pathologic data. Consider this option 1, which accepts that certain FNAB results are considered definitive despite the absence of a gold standard histopathologic result.

What if other ways of studying TIRADS are considered? A full evaluation of TIRADS (consider this Option 2) would require all nodules to be assessed in the same way (i.e., surgical pathology), operating on cytologically benign nodules in patients knowing that the diagnosis of cancer is very low; this is not ethical. Now consider selecting patients based on ACR-TIRADS that would warrant biopsy (TIRADS 3–5) (Option 3), the LR₊s are different as seen in Table 21.6. Based on ACR-TIRADS, TIRADS 1 and 2 nodules, regardless of size, do not require any further biopsy or follow-up.

TIRADS 5 is less likely to change the pretest probability than was the case in option 1, because LR₊ is now 7.15 instead of 11.11, TIRADS 4 is reduced to 0.43 from 0.67, and TIRADS 3 is reduced to 0.059 from 0.092 .

Which should be chosen? Option 2 is not feasible. Option 3 discards the information on TIRADS 1 and 2, even though that is incomplete, and lowers the LR_s. We are left, then, with option 1. See Table 21.3.

The article by Huh et al. [7] used both FNAB (Bethesda II and VI) and tissue pathology to categorize lesions using clinically applicable definitions of benign and malignant. The authors acknowledged the inherent limitations of incorporating cytology as a reference standard and the possibility of discordant histology had surgery been undertaken, however, for Bethesda II and VI this would be exceedingly low. There were 109 Bethesda V nodules, which did not undergo subsequent diagnostic clarification, not clarified in the text but possibly due

to active surveillance or patient preference against further investigations for diagnostic clarification; and these unclarified nodules may have lowered the prevalence of malignant lesions in the study.

Will the Results Help Me in Caring for My Patient?

3(a) Will the reproducibility of the test result and its interpretation be satisfactory in my setting?

Your center already uses the ACR-TIRADS and Bethesda Systems, so application of this test will be straightforward, unlike situations where neither of these systems are currently used.

3(b) Are the results applicable to my patient?

We have assessed the validity of the Huh et al. article [7] and its results, and now want to know if it helps us in caring for the patient. The pretest probability of cancer is estimated at 3%. Let’s say in our clinical scenario that the patient repeated an US at your institution and was found to have a 2 cm TIRADS 4 nodule. This meets the criteria for “positive” based on ACR-TIRADS criteria. Application of the LR of 2.34 gives a posttest probability of 6.7% (see Nomogram in Fig. 21.2). This is not enough to exceed the patient’s threshold for surgery, although the risk of malignancy is not negligible and would be high enough to consider further diagnostic testing (see Fig. 21.3).

You decide, therefore, to obtain an US-guided FNAB and apply LR from that test [17]. The result was read as Bethesda Level IV. Taking the pretest probability as 6.7%, and applying the LR of 1.97, on the Fagan nomogram the posttest probability becomes about 12.4%. The patient is impressed that so much more information could

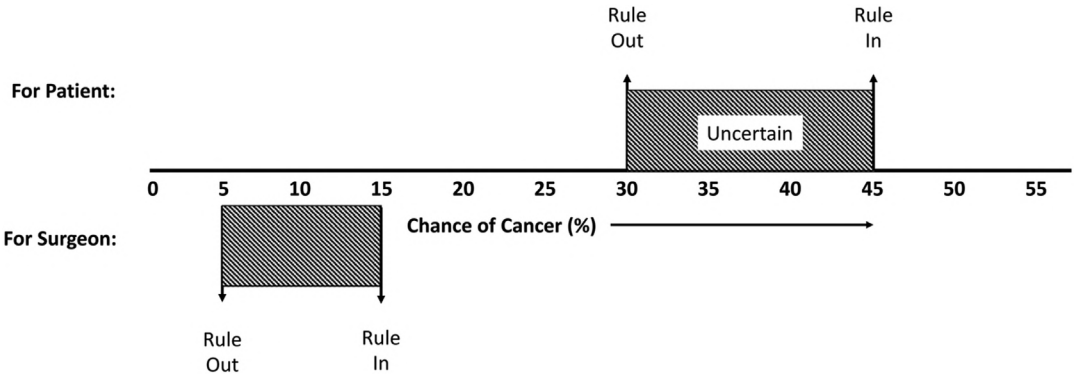


Fig. 21.3 Rule in, rule out line

Table 21.7 ThyroSeq version 3: performed on thyroid nodules subjected to FNAB, and read as “Bethesda III” (follicular neoplasm or suspicious for a follicular neoplasm)

	Malignant	Benign	Totals
“Positive”	32 (TP) ^a <i>a</i> ^b	18 (FP) <i>b</i>	50 <i>a + b</i>
“Negative”	3 (FN) <i>c</i>	101 (TN) <i>d</i>	104 <i>c + d</i>
Totals	35 <i>a + c</i>	119 <i>b + d</i>	154 <i>a + b + c + d</i>

^a*TP* True Positive; *FP* False Positive, *FN* False Negative; *TN* True Negative, denote the “cell names”

^b“*a*,” “*b*,” “*c*,” “*d*” refer to the numerical values of those cells: 32, 18, 3, 101, respectively

be obtained from minimally invasive tests, but she remains unconvinced that surgery is necessary. In the absence of diagnostic surgery, her nodule falls under indeterminate cytology without diagnostic clarification. One could consider repeat biopsy, core biopsy, or diagnostic hemithyroidectomy. Molecular testing is also an option, and she wants to know how the new test, ThyroSeq ver3, might affect the probabilities.

3(c) Will the results change my management?

You search for the diagnostic performance of ThyroSeq ver3 in Bethesda IV nodules and find an article by Steward et al. [20]. Information is available to allow calculation of the diagnostic test variables. Seeing that multiple diagnostic categories of Bethesda Classification are reported is reassuring that the article is compre-

Table 21.8 ThyroSeq version 3: Performed on thyroid nodules subjected to FNAB, and read as “Bethesda IV” (follicular neoplasm or suspicious for a follicular neoplasm)

	Malignant	Benign	Totals
“Positive”	32 (TP) ^a <i>a</i> ^b	15 (FP) <i>b</i>	47 <i>a + b</i>
“Negative”	1 (FN) <i>c</i>	45 (TN) <i>d</i>	46 <i>c + d</i>
Totals	33 <i>a + c</i>	60 <i>b + d</i>	93 <i>a + b + c + d</i>

^a*TP* True Positive; *FP* False Positive, *FN* False Negative; *TN* True Negative, denote the “cell names”

^b“*a*,” “*b*,” “*c*,” “*d*” refer to the numerical values of those cells: 32, 15, 1, 45, respectively

hensive. Two of the categories, III and IV, are shown here, with calculation of LR_s (Tables 21.7 and 21.8).

You use the data in Tables 21.7 and 21.8 to determine LR_s. Specifically, the patient’s FNAB demonstrated Bethesda IV, allowing you to focus on the data in Table 21.8.

$$LR+ = \frac{a / (a + c)}{b / (b + d)} = \frac{32 / 33}{15 / 60} = 3.88$$

This does not exceed 10, so it indicates a modest change in probabilities from pre- to posttest.

$$LR- = \frac{c / (a + c)}{d / (b + d)} = \frac{1 / 33}{45 / 60} = 0.04$$

At 0.04, this rules out the diagnosis of malignancy

Based on the data, ThyroSeq ver3 in the setting of Bethesda IV nodules has a LR+ of 3.88 and LR− of 0.04. You can tell the patient that if the test were applied, and she “fit” the study sample of patients, then using the Fagan nomogram, a positive test would change her pretest probability of cancer from 12.4% to a posttest probability of 35.5%, and if the test is negative, the pretest probability would change from 12.4% to a posttest probability of 0.6%.

You also point out that if the ThyroSeq ver3 test had been applied initially without first doing the US-TIRADS and the FNAB in sequence as was done, given a pretest probability of 3% and an LR+ of 3.88, then the posttest probability would be 10.7%, not high enough for her to choose surgery, although a negative test with LR− of 0.04 would have ruled out cancer, with a posttest probability of less than 0.1%.

3(d) Will patients be better off as a result of the test?

The use of sequential tests for which the LR_s are known or can be calculated as shown in this example have proved to be very helpful. It is important to use all levels available in the data for LR. With multiple levels, there is a greater chance to find the LR > 10, which is sufficient to move almost any pretest probability of disease to the posttest “rule in” disease threshold (Fig. 21.3). Now that you have worked this out for this clinical problem, and you have the LR_s already calculated, and the nomogram available, future patients with the same problem can be approached with confidence and efficiency.

Caution should be used when applying the latest new “wonder test” to try to gain the requisite information all at once, as these tests need to be subjected to the same rigorous scrutiny as was demonstrated for TIRADS to be considered valid and applicable to a particular situation.

Appendix 1: Results of Cochrane Literature Reviewed

1. Lu K, Wang L, Lai S, Chen Z, Cong S, Huang C, Gan K, Guan H, Kuang J. A new discriminant strategy combined with four TIRADS

screening procedures increases ultrasound diagnostic accuracy-focusing on “wrong diagnostic” thyroid nodules. *Eur Radiol.* 2023;33(2):784–96.

2. <https://classic.clinicaltrials.gov/ct2/show/NCT05583097>.
3. Zhu H, Yang Y, Wu S, Chen K, Luo H, Huang J. Diagnostic performance of US-based FNAB criteria of the 2020 Chinese guideline for malignant thyroid nodules: comparison with the 2017 American College of Radiology guideline, the 2015 American Thyroid Association guideline, and the 2016 Korean Thyroid Association guideline. *Quant Imaging Med Surg.* 2021;11(8):3604–18.

Appendix 2: Results from OVID Medline Literature Search

1. Kang S, Kwon SK, Choi HS, Kim MJ, Park YJ, Park DJ, Cho SW. Comparison of Korean vs. American Thyroid Imaging Reporting and Data System in malignancy risk assessment of indeterminate thyroid nodules. *Endocrinol Metab (Seoul, Korea).* 2021;36(5):1111–20.
2. Freire da Silva P, Correa de Araujo Arcoverde L, de Siqueira Barbosa Arcoverde L, Tenorio Wanderley Fernandes Lima G, Paes de Medeiros Lima T, Jose do Amaral F, Bandeira F. Agreement Between American and European Thyroid Imaging, Reporting, and Data System (TIRADS) in the Diagnosis of 473 Thyroid Nodules From a Single Center in Brazil. *Endocr Pract.* 2021;27(11):1108–13.
3. Li C, Wen D, Lu H, Chen Y, Liu X. The value of ACR-TIRADS and C-TIRADS in the diagnosis of nodular Hashimoto thyroiditis and papillary thyroid carcinoma with Hashimoto thyroiditis]. *Lin Chung Er Bi Yan Hou Tou Jing Wai Ke Za Zhi.* 2022;36(6):447–52.
4. Li H-J, Yang Y-P, Liang X, Zhang Z, Xu X-H. Comparison of the diagnostic performance of three ultrasound thyroid nodule risk stratification systems for follicular thyroid neoplasm: K-TIRADS, ACR -TIRADS,

- and C-TIRADS. *Clin Hemorheol Microcirc*. 2023;85(4):395–406.
5. Ha EJ, Lee JH, Lee DH, Na DG, Kim J-H. Development of a machine learning-based fine-grained risk stratification system for thyroid nodules using predefined clinico-radiological features. *Eur Radiol*. 2023;33(5):3211–21.
 6. Chen H, Ye J, Song J, You Y, Chen W, Liu Y. Comparison of different ultrasound classification systems of thyroid nodules for identifying malignant potential: A cross-sectional study. *Clinics (Sao Paulo)*. 2021;76:e2126.
 7. Huang Y, Hong Y, Xu W, Song K, Huang P. Contrast-enhanced ultrasound improves the accuracy of the ACR TI-RADS in the diagnosis of thyroid nodules located in the Isthmus. *Kontrastverstärkter Ultraschall verbessert die Genauigkeit des ACR-TIRADS bei der Diagnose von Schilddrüsenknoten im Isthmus*. 2022;43(6):599–607.
 8. Hekimsoy I, Ozturk E, Ertan Y, Orman MN, Kavukcu G, Ozgen AG, Ozdemir M, Ozbek SS. Diagnostic performance rates of the ACR-TIRADS and EU-TIRADS based on histopathological evidence. *Diagn Interv Radiol*. 2021;27(4):511–8.
 9. Sakthisankari S, Vidhyalakshmi S, Shanthakumari S, Devanand B, Nagul U. The combination of ACR-Thyroid Imaging Reporting and Data system and The Bethesda System for Reporting Thyroid Cytopathology in the evaluation of thyroid nodules-An institutional experience. *Cytopathology*. 2021;32(4):472–81.
 10. Huh S, Yoon JH, Lee HS, Moon HJ, Park VY, Kwak JY. Comparison of diagnostic performance of the ACR and Kwak TIRADS applying the ACR TIRADS' size thresholds for FNA. *Eur Radiol*. 2021;31(7):5243–5250.
 11. Cui Y, Fu C, Si C, Li J, Kang Y, Huang Y, Cui K. Analysis and comparison of the malignant thyroid nodules not recommended for biopsy in ACR TIRADS and AI TIRADS with a large sample of surgical series. *J Ultrasound Med*. 2023;42(6):1225–33.
 12. Kim DH, Kim SW, Basurrah MA, Lee J, Hwang SH. Diagnostic performance of six ultrasound risk stratification systems for thyroid nodules: a systematic review and network meta-analysis. *AJR Am J Roentgenol*. 2023;220(6):791–803.
 13. Yucel S, Balci IG, Tomak L. Diagnostic performance of thyroid nodule risk stratification systems: comparison of ACR-TIRADS, EU-TIRADS, K-TIRADS, and ATA guidelines. *Ultrasound Q*. 2023;39(4):206–11.
 14. Qi Q, Zhou A, Guo S, Huang X, Chen S, Li Y, Xu P. Explore the diagnostic efficiency of chinese thyroid imaging reporting and data systems by comparing with the other four systems (ACR TI-RADS, kwak-TIRADS, KSThR-TIRADS, and EU-TIRADS): A single-center study. *Front Endocrinol (Lausanne)*. 2021;12:763897.
 15. Chambara N, Liu SYW, Lo X, Ying M. Diagnostic performance evaluation of different TI-RADS using ultrasound computer-aided diagnosis of thyroid nodules: An experience with adjusted settings. *PLoS One*. 2021;16(1):e0245617.
 16. Marina M, Zatelli MC, Goldoni M, Del Rio P, Corcione L, Martorana D, Percesepe A, Bonatti F, Mozzoni P, Crociara A, Ceresini G. Combination of ultrasound and molecular testing in malignancy risk estimate of Bethesda category IV thyroid nodules: results from a single-institution prospective study. *J Endocrinol Invest*. 2021;44(12):2635–43.
 17. Tuli G, Munarin J, Scollo M, Quagliano F, De Sanctis L. Evaluation of the efficacy of EU-TIRADS and ACR-TIRADS in risk stratification of pediatric patients with thyroid nodules. *Front Endocrinol (Lausanne)*. 2022;13:1041464.
 18. Xing Z, Qiu Y, Zhu J, Su A, Wu W. Diagnostic performance of adult-based ultrasound risk stratification systems in pediatric thyroid nodules: a systematic review and meta-analysis. *Front Endocrinol (Lausanne)*. 2023;14:1187935.
 19. Petersen M, Schenke SA, Seifert P, Stahl AR, Gorges R, Grunert M, Klemenz B, Kreissl MC, Zimny M. Correct and incorrect recom-

- mendations for or against fine needle biopsies of hypofunctioning thyroid nodules: performance of different ultrasound-based risk stratification systems. *Richtige und falsche Empfehlungen für oder gegen Feinnadelbiopsien von hypofunktionellen Schilddrüsenknoten Leistung Verschied Ultraschall-Risikostratifizierungssysteme.* 2024;63(1):21–33.
20. Joo L, Lee MK, Lee JY, Ha EJ, Na DG. Diagnostic performance of ultrasound-based risk stratification systems for thyroid nodules: a systematic review and meta-analysis. *Endocrinol Metab (Seoul, Korea).* 2023;38(1):117–28.
 21. Chatti H-A, Oueslati I, Azaiez A, Marrakchi J, Boukriba S, Mizouni H, Haouet S, Besbes G, Yazidi M, Chihouai M. Diagnostic performance of the EU TI-RADS and ACR TI-RADS scoring systems in predicting thyroid malignancy. *Endocrinol diabetes Metab.* 2023;6(4):e434.
 22. Li W, Wang Y, Wen J, Zhang L, Sun Y. Diagnostic performance of American College of Radiology TI-RADS: a systematic review and meta-analysis. *AJR Am J Roentgenol.* 2021;216(1):38–47.
 23. Lin Y, Luo Y-K, Li J, Ren X-Y, Wang H-W. Diagnostic Value of American Thyroid Association Guidelines, American College of Radiology Thyroid Imaging Reporting and Data System, and Chinese Thyroid Imaging Reporting and Data System Alone and Combined With BRAFV600E Mutation in Atypia of Undetermined. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao.* 2023;45(6):921–8.
 24. Yang L, Li C, Chen Z, He S, Wang Z, Liu J. Diagnostic efficiency among Eu-/C-/ACR-TIRADS and S-Detect for thyroid nodules: a systematic review and network meta-analysis. *Front Endocrinol (Lausanne).* 2023;14:1227339.
 25. Hawkins SP, Jamieson SG, Coomarasamy CN, Low IC. The global epidemic of thyroid cancer overdiagnosis illustrated using 18 months of consecutive nodule biopsy correlating clinical priority, ACR-TIRADS and Bethesda scoring. *J Med Imaging Radiat Oncol.* 2021;65(3):309–16.
 26. Han D, Ibrahim N, Lu F, Zhu Y, Du H, AlZoubi A. Automatic detection of thyroid nodule characteristics from 2D ultrasound images. *Ultrason Imaging.* 2024;46(1):41–55.
 27. Zheng T, Xie X, Ni Z, Tang L, Wu P-Y, Song B. Quantitative evaluation of diffusion-weighted MRI for differentiating benign and malignant thyroid nodules larger than 4 cm. *BMC Med Imaging.* 2023;23(1):212.
 28. Kim PH, Yoon HM, Hwang J, Lee JS, Jung AY, Cho YA, Baek JH. Diagnostic performance of adult-based ATA and ACR-TIRADS ultrasound risk stratification systems in pediatric thyroid nodules: a systematic review and meta-analysis. *Eur Radiol.* 2021;31(10):7450–7463.
 29. Xu S, Ni X, Zhou W, Zhan W, Zhang H. Development and validation of a novel diagnostic tool for predicting the malignancy probability of thyroid nodules: A retrospective study based on clinical, B-mode, color doppler and elastographic ultrasonographic characteristics. *Front Endocrinol (Lausanne).* 2022;13:966572.
 30. Li Q, Yang L, Yang L, Jiang X, Li S. Utility of six ultrasound-based risk stratification systems in the diagnosis of AUS/FLUS thyroid nodules. *Acad Radiol.* 2024;31(1):131–41.
 31. Chen Q, Lin M, Wu S. Validating and Comparing C-TIRADS, K-TIRADS and ACR-TIRADS in stratifying the malignancy risk of thyroid nodules. *Front Endocrinol (Lausanne).* 2022;13:899575.
 32. Jin Z, Pei S, Shen H, Ouyang L, Zhang L, Mo X, Chen Q, You J, Zhang S, Zhang B. Comparative study of C-TIRADS, ACR-TIRADS, and EU-TIRADS for diagnosis and management of thyroid nodules. *Acad Radiol.* 2023;30(10):2181–91.
 33. Liu J, Luo T, Zhang H, Liu H, Gu Y, Chen X, Shi L, Guan L, Ni X, Zhang X, Zhang R, Jia X, Dong Y, Zhang J, Xu W, Zhou J. Markedly hypoechoic: a new definition improves the

- diagnostic performance of thyroid ultrasound. *Eur Radiol.* 2023;33(11):7857–65.
34. Behbahaninia M, Nabahati M, Moazezi Z, Mehraeen R, Shirafkan H. Diagnostic performance of ACR-TIRADS, Korean TIRADS, and American Thyroid Association guidelines for risk stratification of thyroid nodules: a prospective study. *J Ultrasound.* 2022;25(4):887–94.
 35. Kim PH, Suh CH, Baek JH, Chung SR, Choi YJ, Lee JH. Unnecessary thyroid nodule biopsy rates under four ultrasound risk stratification systems: a systematic review and meta-analysis. *Eur Radiol.* 2021;31(5):2877–85.
 36. Lee YJ, Cho YJ, Heo YJ, Chung E-J, Choi YH, Kim J-I, Park YJ, Shin CH, Lee YA. Thyroid nodules in childhood-onset Hashimoto's thyroiditis: Frequency, risk factors, follow-up course and genetic alterations of thyroid cancer. *Clin Endocrinol (Oxf).* 2021;95(4):638–48.
 37. Hoang JK, Middleton WD, Langer JE, Schmidt K, Gillis LB, Nair SS, Watts JA, Snyder RW 3rd, Khot R, Rawal U, Tessler FN. Comparison of thyroid risk categorization systems and fine-needle aspiration recommendations in a multi-institutional thyroid ultrasound registry. *J Am Coll Radiol.* 2021;18(12):1605–13.
 38. Araruna Bezerra de Melo R, Menis F, Calsavara VF, Stefanini FS, Novaes T, Saieg M. The impact of the use of the ACR-TIRADS as a screening tool for thyroid nodules in a cancer center. *Diagn Cytopathol.* 2022;50(1):18–23.
 39. Bas H, Ustuner E, Kula S, Konca C, Demirel S, Elhan AH. Elastography and doppler may bring a new perspective to TIRADS, altering conventional ultrasonography dominance. *Acad Radiol.* 2022;29(3):e25–38.
 40. Hoang JK, Asadollahi S, Durante C, Hegedus L, Papini E, Tessler FN. An international survey on utilization of five thyroid nodule risk stratification systems: a needs assessment with future implications. *Thyroid.* 2022;32(6):675–81.
 41. Chen X, Kutaiba N, Pearce S, Digby S, Van Gelderen D. Application of Thyroid Imaging Reporting and Data System (TIRADS) guidelines to thyroid nodules with cytopathological correlation and impact on healthcare costs. *Intern Med J.* 2022;52(8):1366–73.
 42. Babajani A, Rahmani S, Raoufi M, Eidgahi ES, Dastjerdi AV, Behfarnia P, Khalili S, Moghaddam NA. Clinico-cytopathological subcategorization in thyroid nodules of atypia of undetermined significance/follicular lesion of undetermined significance using the TIRADS and Bethesda classifications. *Front Endocrinol (Lausanne).* 2023;14:1135196.
 43. de Sanctis L, Wasniewska M, Vigone MC. Editorial: Thyroid nodules and tumors in childhood. *Front Endocrinol (Lausanne).* 2023;14:1303749.
 44. Malhi HS, Grant EG. Ultrasound of thyroid nodules and the thyroid imaging reporting and data system. *Neuroimaging Clin N Am.* 2021;31(3):285–300.
 45. Gonzalez Vasquez CM, Munoz Duran JA, Isaza Zapata S, Gonzalez Londono JF, Garcia Gomez V. Concordance of the ACR TIRADS. *Radiologia.* 2021;63(6):469–475.
 46. Wu M-H, Chen K-Y, Hsieh M-S, Chen A, Chen C-N. Risk stratification in patients with follicular neoplasm on cytology: use of quantitative characteristics and sonographic patterns. *Front Endocrinol (Lausanne).* 2021;12:614630.
 47. Eidt LB, Nunes de Oliveira C, Lagos YBB De, Solera GLM, Izquierdo R, Meyer EL de S, Mattevi VS, Golbert L. A prospective comparison of ACR-TIRADS and EU-TIRADS in thyroid nodule assessment for FNA-US. *Clin Endocrinol (Oxf).* 2023;98(3):415–25.
 48. McClean S, Omakobia E, England RJA. Comparing ultrasound assessment of thyroid nodules using BTA U classification and ACR TIRADS measured against histopathological diagnosis. *Clin Otolaryngol.* 2021;46(6):1286–9.

49. Mao S, Zhao LP, Li XH, Sun YF, Su H, Zhang Y, Li KL, Fan DC, Zhang MY, Sun ZG, Wang SC. [The diagnostic performance of 2020 Chinese Ultrasound Thyroid Imaging Reporting and Data System in thyroid nodules]. *Zhonghua Yi Xue Za Zhi*. 2021;101(45):3748–53.
50. Lin Y, Lai S, Wang P, Li J, Chen Z, Wang L, Guan H, Kuang J. Performance of current ultrasound-based malignancy risk stratification systems for thyroid nodules in patients with follicular neoplasms. *Eur Radiol*. 2022;32(6):3617–30.
51. Zhou L, Zheng L-L, Zhang C-J, Wei H-F, Xu L-L, Zhang M-R, Li Q, He G-F, Ghamor-Amegavi EP, Li S-Y. Comparison of S-Detect and thyroid imaging reporting and data system classifications in the diagnosis of cytologically indeterminate thyroid nodules. *Front Endocrinol (Lausanne)*. 2023;14:1098031.
52. Cai M, Chen L, Shui L, Lv X, Wang H. Explore the diagnostic performance of 2020 Chinese Thyroid Imaging Reporting and Data Systems by comparing with the 2017 ACR-TIRADS guidelines: a single-center study. *Endocrine*. 2023;80(2):399–407.
53. Bolland MJ, Grey A. Increased workload without clinical benefit: Results following implementation of the ACR-TIRADS system for thyroid nodules. *Clin Endocrinol (Oxf)*. 2023;99(3):328–34.
54. Triggiani V, Lisco G, Renzulli G, Frasoldati A, Guglielmi R, Garber J, Papini E. The TNAPP web-based algorithm improves thyroid nodule management in clinical practice: A retrospective validation study. *Front Endocrinol (Lausanne)*. 2022;13:1080159.
55. Dhar L, Singh M, Jain SL, Rana D, Kumar J, Meher R, Khurana N, Verma N, Mohammed A, Chellani G. Correlation of ultrasound-based TIRADS and the Bethesda system for reporting of thyroid cytopathology: A study in a tertiary care centre. *Cytopathology*. 2023;34(6):590–6.
56. Orhan Soylemez UP, Gunduz N. Diagnostic accuracy of five different classification systems for thyroid nodules: A prospective, comparative study. *J Ultrasound Med*. 2022;41(5):1125–36.
57. Trimboli P, Castellana M, Piccardo A, Romanelli F, Grani G, Giovannella L, Durante C. The ultrasound risk stratification systems for thyroid nodule have been evaluated against papillary carcinoma. A meta-analysis. *Rev Endocr Metab Disord*. 2021;22(2):453–460.
58. Staibano P, Forner D, Noel CW, Zhang H, Gupta M, Monteiro E, Sawka AM, Pasternak JD, Goldstein DP, de Almeida JR. Ultrasonography and fine-needle aspiration in indeterminate thyroid nodules: a systematic review of diagnostic test accuracy. *Laryngoscope*. 2022;132(1):242–51.
59. Sakajiri RK, Rahal Junior A, Francisco Neto MJ, Queiroz MRG de, Garcia RG, Martins LAL, Malerbi DAC. Ultrasound classification of thyroid nodules: does size matter?. *Einstein (Sao Paulo)*. 2022;20:eAO6747.
60. Ma Y, Huo X, Kong S, Xu W, Zhao W, Zhu M. A Review about C-TIRADS, ACR-TIRADS, and K-TIRADS Combined with Real-Time Tissue Elastography to Diagnose Thyroid Nodules. *Discov Med*. 2023;35(174):1–10.
61. Mena G, Montalvo A, Ubidia M, Olmedo J, Guerrero A, Leon-Rojas JE. Elastography of the thyroid nodule, cut-off points between benign and malignant lesions for strain, 2D shear wave real time and point shear wave: a correlation with pathology, ACR TIRADS and Alpha Score. *Front Endocrinol (Lausanne)*. 2023;14:1182557.
62. Lu K, Wang L, Lai S, Chen Z, Cong S, Huang C, Gan K, Guan H, Kuang J. A new discriminant strategy combined with four TIRADS screening procedures increases ultrasound diagnostic accuracy-focusing on “wrong diagnostic” thyroid nodules. *Eur Radiol*. 2023;33(2):784–96.
63. Burgos N, Zhao J, Brito JP, Hoang JK, Pitoia F, Maraka S, Castro MR, Lee J-H, Singh Ospina N. Clinician agreement on the classification of thyroid nodules ultrasound fea-

tures: a survey of 2 endocrine societies. *J Clin Endocrinol Metab.* 2022;107(8):e3288–94.

References

1. Davies L, Welch HG. Current thyroid cancer trends in the United States. *JAMA Otolaryngol Head Neck Surg.* 2014;140(4):317–22.
2. Topstad D, Dickinson JA. Thyroid cancer incidence in Canada: a national cancer registry analysis. *CMAJ Open.* 2017;5(3):E612–6.
3. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991. p. 95.
4. Corso C, Gomez X, Sanabria A, Vega V, Dominguez LC, Osorio C. Total thyroidectomy versus hemithyroidectomy for patients with follicular neoplasm. A cost-utility analysis. *Int J Surg.* 2014;12:837–42.
5. Cooper DS, Doherty GM, Haugen BR, Kloos RT, Lee SL, Mandel SJ, Mazzaferri EL, McIver B, Pacini F, Schlumberger M, Sherman SI, Steward DL, Tuttle RM. For the American Thyroid Association (ATA) guidelines taskforce on thyroid nodules and differentiated thyroid cancer. Revised American thyroid association management guidelines for patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association (ATA) guidelines taskforce on thyroid nodules and differentiated thyroid cancer. *Thyroid.* 2009;19:1167–214. Errata in: *Thyroid.* 2010;20:674–5; *Thyroid.* 2010;20:942.
6. Welker MJ, Orlov D. Thyroid nodules. *Am Fam Phys.* 2003;67:559–66.
7. Huh S, Yoon JH, Lee HS, Moon HJ, Park VY, Kwak JY. Comparison of diagnostic performance of the ACR and Kwak TIRADS applying the ACR TIRADS' size thresholds for FNA. *Eur Radiol.* 2021;31(7):5243–50.
8. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, Schuff KG, Sherman SI, Sosa JA, Steward DL, Tuttle RM, Wartofsky L. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. *Thyroid.* 2016;26(1):1–133.
9. Kim DH, Kim SW, Basurrah MA, Lee J, Hwang SH. Diagnostic performance of six ultrasound risk stratification systems for thyroid nodules: a systematic review and network meta-analysis. *Am J Roentgenol.* 2023;220(6):791–803.
10. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, Cronan JJ, Beland MD, Desser TS, Frates MC, Hammers LW, Hamper UM, Langer JE, Reading CC, Scoutt LM, Stavros AT. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. *J Am Coll Radiol.* 2017;14(5):587–95.
11. Archibald S, Bhandari M, Thoma A. For the evidence-based surgery working group. Users' guides to the surgical literature: how to use an article about a diagnostic test. *CJS.* 2001;44(1):17–23.
12. Fagan TJ. A nomogram for applying likelihood ratios [letter]. *NEJM.* 1975;293:257.
13. Parikh R, Parikh S, Arun E, Thomas R. Likelihood ratios: clinical application in day-to-day practice. *Indian J Ophthalmol.* 2009;57(3):217–21.
14. Kwak JY, Han KH, Yoon JH, Moon HJ, Son EJ, Park SH, Jung HK, Choi JS, Kim BM, Kim EK. Thyroid imaging reporting and data system for US features of nodules: A step in establishing better stratification of cancer risk. *Radiology.* 2011;260(3):892–9.
15. Horvath E, Majlis S, Rossi R, Franco C, Niedmann JP, Castro A, Dominguez M. An ultrasonogram reporting system for thyroid nodules stratifying cancer risk. *J Clin Endocrinol Metab.* 2009;94:1748–51.
16. Middleton WD, Teefey SA, Reading CC, Langer JE, Beland MD, Szabunio MM, Desser TS. Multiinstitutional analysis of thyroid nodule risk stratification using the American College of Radiology Thyroid Imaging Reporting and Data System. *AJR Am J Roentgenol.* 2017;208(6):1331–41. <https://doi.org/10.2214/AJR.16.17613>.
17. Dhyani M, Faquin W, Lubitz CC, Daniels GH, Samir AE. How to interpret thyroid fine-needle aspiration biopsy reports: a guide for the busy radiologist in the era of the Bethesda Classification system. *AJF.* 2013;201(6):1335–9.
18. Bongiovanni M, Spitale A, Faquin WC, Mazzucchelli L, Baloch ZW. The Bethesda system for reporting thyroid cytopathology: a meta-analysis. *Acta Cytol.* 2012;56:333–9.
19. Jaeschke R, Guyatt G, Sackett DL. The evidence-based medicine working group. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A: are the results of the study valid? *JAMA.* 1994;271(9):703–7.
20. Steward DL, Carty SE, Sippel RS, Yang SP, Sosa JA, Sipos JA, Figge JJ, Mandel S, Haugen BR, Burman KD, Baloch ZW, Lloyd RV, Seethala RR, Gooding WE, Chiosea SI, Gomes-Lima C, Ferris RL, Folek JM, Khawaja RA, Kundra P, Loh KS, Marshall CB, Mayson S, McCoy KL, Nga ME, Ngiam KY, Nikiforova MN, Poehls JL, Ringel MD, Yang H, Yip L, Nikiforov YE. Clinical validation of THYROSEQ V3 performance in thyroid nodules with indeterminate cytology: a prospective blinded multi-institutional validation study. *Thyroid.* 2017;27(Suppl1):A–167.

How to Assess a Prognostic Study

22

Saurabh Gupta, Kevin Kim, Emilie Belley-Côté,
and Richard P. Whitlock

Abbreviations

AS	Aortic Stenosis
AVIATOR	Aortic Valve Insufficiency and Ascending Aorta Aneurysm International Registry
RCTs	Randomized Controlled Trials
TAVR	Transcatheter Aortic Valve Replacement

S. Gupta
Division of Cardiac Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada

Department of Health Research Methods,
Evidence, and Impact, McMaster University,
Hamilton, ON, Canada

K. Kim
Department of Health Research Methods,
Evidence, and Impact, McMaster University,
Hamilton, ON, Canada

Population Health Research Institute,
Hamilton, ON, Canada

Clinical Scenario

You are a cardiac surgeon consulting on a 45-year-old man with severe aortic insufficiency. He has a bicuspid aortic valve, followed by his cardiologist for years, and is becoming increasingly dyspneic. On physical examination, he weighs 100 kg (220 lbs) and is 183 cm (6 ft) tall, and you auscultate a 3/6 diastolic ejection murmur. His echocardiogram shows a dilated aortic root, a central regurgitation jet, and a left ventricular ejection fraction of 58%. You discuss aortic valve repair instead of replacement. You review the risks and benefits of the surgery. He

E. Belley-Côté
Department of Health Research Methods,
Evidence, and Impact, McMaster University,
Hamilton, ON, Canada

Population Health Research Institute,
Hamilton, ON, Canada

Division of Cardiology & Critical Care, Department
of Medicine, McMaster University,
Hamilton, ON, Canada
e-mail: emilie.belley-cote@phri.ca

R. P. Whitlock (✉)
Department of Surgery, Division of Cardiac Surgery,
McMaster University, Hamilton, ON, Canada

Department of Health Research Methods,
Evidence, and Impact, McMaster University,
Hamilton, ON, Canada

Population Health Research Institute,
Hamilton, ON, Canada
e-mail: richard.Whitlock@phri.ca

asks you two questions: “How long will the repair last?” and “What are the possible operative and long-term complications?” You ask your senior resident to check for prognostic studies around bicuspid aortic valve repair surgery to discuss during the patient’s next appointment.

How and Why We Measure Prognosis?

As surgeons, we aim for interventions that do more benefit than harm and intend for better outcomes than the natural course of the pathology. Prognostic studies examine the effect of variables, like surgical interventions, on disease progression [1]. Understanding prognosis can help us better advise our patients on treatment options and natural disease progression. Further, knowledge of expected outcomes after a given procedure informs us, and our patients, in the decision-making process [2].

Providing accurate estimates of patient prognosis requires a study that reports outcomes in patients with similar clinical presentation [3]. For example, a surgeon looking to determine in-hospital mortality after replacing the ascending aorta and aortic valve in a patient with Marfan syndrome would need studies that report outcomes in other Marfan patients. Such patient data are referred to as prognostic factors, clinical characteristics that are objectively measured and can help predict patient outcomes [4].

In the remainder of this chapter, we will focus on identifying trustworthy prognostic information using the results in patient care.

Study Designs for Prognostic Studies

Before undertaking a literature search, it is important to understand what study designs best answer our clinical question. When determining the effectiveness of treatment, we consider randomized controlled trials (RCTs) as high quality evidence and observational studies as lower quality [5]. When determining prognosis, however,

we might place more confidence in estimates of prognosis from observational studies than RCTs [6]. This is because RCTs include filters in their eligibility criteria (e.g., age restriction, comorbidity, drug intolerance, etc.) that exclude patients relevant to the broader prognostic question of interest. Eligible patients may decline to participate in an RCT due to reasons related to their prognosis. One exception to this rule are large, simple, pragmatic trials with broad eligibility criteria. For example, Stassano et al. [7] randomized 310 patients to receive a mechanical or bioprosthetic aortic valve. If our patient wanted to compare effectiveness of one valve type over the other, “Aortic valve replacement: a prospective randomized evaluation of mechanical versus biological valves in patients ages 55 to 70 years” by Stassano et al. [7] might be a good study to use. However, our patient wants to know his prognosis specifically after an aortic valve repair with or without a valve-sparing root replacement. As such, a cohort study is ideal in identifying associations between our prognostic factors of interest and outcomes [7].

In cohort studies (a type of observational study), patients receive the intervention (i.e., they are not randomized) and are observed over a period. Furthermore, cohort studies can be conducted prospectively—forward in time—or retrospectively—data are collected on patients who received the intervention some time ago, and we want to know their outcome [8] (See Chap. 17). Both approaches have advantages and disadvantages. Prospective designs are powerful for assessing incidence and investigating potential causes. They measure the stage of disease prior to treatment, preventing the true effect from being influenced by the knowledge of the outcome. Most importantly, investigators using a prospective cohort design can measure outcomes more completely and accurately than a retrospective design [9]. The disadvantages are high cost and inefficiency when studying rare outcomes. Alternatively, a retrospective study design is less costly and time-intensive. A retrospective cohort study fundamentally has the same methodology as a prospective study, except that it looks back in time. Subjects

already had their exposure to the variable of interest, had baseline measurements recorded, and follow-up period completed. The main disadvantage of a retrospective study is that biases may affect selection of patients and recall of information. Further, we must use the data available, which may be incomplete, inaccurate, or measured in ways not suited to answer our question [10]. The 2021 study by Jasinski et al., which we chose to help answer our patient’s questions, is a retrospective cohort study. The authors identified patients who underwent aortic valve repair from 2003 through 2020 and evaluated their outcomes [11].

Investigators can also use a “case-control” design to study prognosis. They evaluate patients who have experienced an outcome of interest (cases) and compare them to those without the outcome (controls) [12]. Investigators are then able to assess the relative frequency of prognostic factors among both study groups. For example, Sleder et al. [14] retrospectively evaluated socioeconomic and racial disparities among severe aortic stenosis (AS) patients receiving transcatheter aortic valve replacement (TAVR) in a case-control study. They compared 67 patients with severe AS who underwent TAVR (cases) to patients with severe AS who were not offered TAVR (controls). They noted a statistically significant income disparity between the “cases” and “controls” [13]. Case-control studies are retrospective and share the limitations of retrospective cohort studies. In addition, case-control studies provide relative odds of outcomes, but not absolute risks. Case-control studies are particularly useful when the outcome of interest is rare or requires lengthy time to occur—beyond what would be feasible for a prospective study design [14] (See Chap. 18).

Literature Search

To identify relevant literature to our patient with aortic stenosis, we entered relevant keywords into an established database of publications (Please see Chap. 4). In our scenario, entering the keywords “bicuspid” AND “aortic

valve” AND “repair,” along with activating filters for human, adults (18 years of age or older), full texts, English language, and published within the last 3 years (2021–2023) identified 59 articles in PubMed. If an initial search does not produce articles of interest, we could employ other strategies such as querying multiple databases (EMBASE and MEDLINE). We can also review references of relevant publications and recent textbooks (hand-searching) or reach out to field experts for guidance. Scanning through the most recent publications, we did not identify any relevant RCTs or prospective cohort study. Certain studies were excluded for focusing on basic science, and others for studying outcomes with specific repair types. The search revealed one promising article: “Bicuspid aortic valve repair: outcomes after 17 years” by Jasinski et al. [11] This is a retrospective cohort study that identified patients who underwent bicuspid aortic valve repair from 2003 through 2020 and evaluated their outcomes. This study seems promising in answering our patient’s questions. The steps in appraising a surgical study that deals with prognosis are shown in Table 22.1.

Table 22.1 User’s guide to surgical literature: Guide to an article about prognosis

I. <i>Guide for validity (study methodology)</i>	
Step 1:	
Is the sample representative?	
Were patients homogeneous with respect to their prognostic risk? Are they at a similar point in the course of the disease?	
If subgroups were identified, did investigators provide estimates for clinically relevant subgroups and adjust for important prognostic factors?	
Step 2:	
Was follow-up sufficiently long and complete?	
Were assessed outcomes objective and unbiased?	
II. <i>Understanding the results</i>	
Step 3: How likely are the outcomes to occur over time?	
Step 4: How precise are the estimates of likelihood?	
III. <i>Using the results to determine patient care (applying the results to your patients)</i>	
Step 5: Can I apply the results to my practice or patient?	

Appraising the Surgical Literature on Prognosis

Before applying the study results to our patient, we must critically appraise the Jasinski et al. publication. Key questions to ask are: (1) Is the sample representative of my patient? (2) Was the follow-up sufficiently long and complete? (3) Were the assessed outcomes objective and unbiased? (4) How likely are the outcomes to occur over time? (5) How precise are the estimates of likelihood? (6) Can I apply the results to my practice or patient? Table 22.1 outlines these questions, and more, that are important when interpreting and using prognostic studies [16].

Is the Sample Representative?

Ideally, we would like to determine prognosis by studying an affected population from onset of exposure to the end. However, we cannot study entire populations and must instead determine prognosis of a sample. As such, evaluating a representative sample of the entire population is crucial for generalizable results [15]. If a sample population is systematically different from the population of interest, it is not representative and may be biased. Biased prognostic studies can result in over- or underestimating event rates. A common way to identify an unrepresentative sample is to look for any systematic processes that patients passed through before entering the study, which may introduce bias; for example, patients with more complex or severe forms of disease are more likely to be referred for tertiary care [3]. Consequently, the prognosis in a tertiary center sample may differ from the prognosis of the disease in general due to referral bias. To determine whether a sample is representative, look for a description of inclusion and exclusion criteria as well as recruitment strategy. In addition, the authors should report objective criteria for sample selection and disease diagnosis. Unclear or inadequate definition of patients within a study increases the risk that the sample will not be representative, and its results biased. This is achieved through precise and consistent

inclusion criteria [16]. For instance, we want studies reporting outcomes in a study group with similar baseline characteristics as our patient from the clinical scenario (45-year-old male with severe, symptomatic aortic insufficiency from a bicuspid aortic valve). If a study included a diverse spectrum of patients undergoing aortic valve repair (a wide age range, for example), a subgroup analysis of study participants like our patients would suffice. In addition to similar disease severity, other prognostic factors such as age, smoking history, diabetes, among others, should be considered. It is important to consider these prognostic factors in relation to each other to avoid making false conclusions.

In situations where a patient characteristic impacts prognosis, a stratified analysis may be used. In a stratified analysis, participants are separated into groups (i.e., strata) by prognostic factors [17]. When many variables predict prognosis, stratified analyses are impractical and statistical methods, like regression, adjust for multiple variables, and determine the strongest prognostic factors [18].

When evaluating the sample in the Jasinski et al. [11] study, we see that the authors obtained patient data from patient medical records from two surgical centers—Medical University of Silesia and Wroclaw Medical University, and echocardiographic data from the Aortic Valve Insufficiency and ascending aorta Aneurysm InternATIOnal Registry (AVIATOR). The authors also used Polish National Health Services and Polish National Registry of Cardiac Surgery Procedures for further details and to ascertain freedom from mortality and reoperation. Mortality outcomes were also analyzed and reported [11]. We can confidently say that Jasinski et al. included a representative sample, with characteristics like our patient's.

Was Follow-Up Sufficiently Long and Complete?

Investigators should follow their patients for an adequate length of time to capture all patient-important outcomes. This is especially important

in determining long-term outcomes, where the risk of loss to follow-up is increased [19].

Losing participants to follow-up threatens the validity of prognostic studies and increases the risk of bias as outcomes in those lost to follow-up may differ from those who remain in the study [19]. For example, anticoagulation therapy is a lifelong commitment for patients receiving a mechanical aortic valve [20]. In a study reporting survival of patients after valve replacement surgery, those lost to follow-up may not be deceased but may have moved or followed up by another physician. The patients surviving and not lost to follow-up are no longer representative of the sample or even the entire study group. The threat to study validity is greatest when outcomes of interest are infrequent, and/or many participants are lost to follow-up. Let us consider the example of mechanical valve thrombosis. In a study by Cannegeiter et al. [23], the incidence of mechanical valve thrombosis in patients receiving appropriate anticoagulant therapy was 0.2 per 100 patient-years. If data for ten participants were lost to follow-up, and they all experienced valve thrombosis, the true incidence of mechanical valve thrombosis would be higher than reported. This carries important clinical implications, suggesting that valve thrombosis was underreported due to loss of follow-up. One of the weaknesses in the Jasinski et al. study is that while they report that mean follow-up for patients was 5 ± 3.5 years (with a range of 17 years), they do not report number lost to follow-up. In consequence, we cannot exclude the possibility of differential loss to follow-up—when the dropout rate differs between treatment groups—potentially introducing bias into the results. Therefore, we must view their data on long-term mortality with caution as the authors do not report completeness of follow-up [21].

Were Objective and Unbiased Outcome Criteria Used?

Outcomes can be objective (e.g., mortality), require adjudication (e.g., myocardial infarction), or be subjective (e.g., quality of life). Whenever possible, investigators should specify outcomes

based on a consensus definition [22]. For instance, a patient's heart failure symptoms can be measured using the New York Heart Association Functional Classification, whereas general quality of life is measured by validated instruments [23]. In the study by Jasinski et al., the authors present mortality and freedom from reoperation as primary outcomes, but also discuss recurrent aortic insufficiency, echocardiographic parameters of LV remodeling, and risk factors for repair failure. While mortality and reoperation are objective outcomes, recurrent aortic insufficiency and "acceptable LV remodelling" requires definition. The authors appropriately define their outcomes in the Material and Methods section.

How Likely Are the Outcomes to Occur over Time?

Results from prognostic studies help determine the risk of events occurring over time [7]. If our patient wanted to understand his 5-year freedom from re-intervention after the repair, we would look for a study that reports outcomes of patients 5 years after their initial surgery. Jasinski et al. demonstrate that in patients with a mean age of 44.5, the probability of freedom from reoperation at 5 years is approximately 0.9 [11]. If our patient asked about changes in his chances of survival over time, we would show him a survival curve, which illustrates the incidence of events over time. To create a survival curve, events must be discrete (death, reoperation, hospitalizations), and the precise time at which they occur must be recorded [24]. According to Jasinski et al., probability of freedom from reoperation at 17 years is approximately 0.8, suggesting a 10% increase in risk between 5 years and 17 years after bicuspid valve repair [11].

How Precise Are the Estimates of Likelihood?

The more precise a prognostic estimate, the closer to certainty and useful it is. Usually, risks of adverse outcomes are reported with their associated 95%

confidence interval (CI), which define the interval within which the true risk likely lies [25]. For example, in a prospective cohort of 196 patients who received bioprosthetic aortic valve replacement, Messe et al. [26] reported clinical strokes in 34 patients (17%, 95% CI [12% to 23%]). That is, assuming the study is valid, there is a 95% chance that the population risk for stroke after the surgery lies between 12% and 23%. Put another way, if the study were repeated 100 times, the incidence of stroke would be between 12% and 23%, approximately 95 of those 100 times. In the Jasinski et al. paper, the authors report 7-year freedom from reoperation as 91.8% (95% CI [0.86 to 0.97]) [11]. This implies that for patients with a mean age of 45.5-years who undergo bicuspid aortic valve repair, we are 95% certain that the risk of reoperation lies between 86 and 97%. For more information on confidence intervals (please see Chap. 28).

Can I Apply the Results to My Practice or Patient?

The authors should describe patient characteristics and demographics in enough detail that we can compare to our own patients and determine applicability. Further, surgical interventions change and evolve over time [27]. The surgical procedure and risks involved with aortic valve repair or replacement in 1975 are not the same in 2018. With evolving valve types and cardiopulmonary bypass technology, outcomes have improved dramatically [28]. Therefore, it is prudent that the study we review be current and applicable to current practice. Jasinski et al. recognized the differences in surgical techniques and bicuspid aortic valve geometry. They reported patient enrollment by specific types of bicuspid aortic valve. These characteristics of the study sample, along with use of these characteristics to determine risk factors for repair failure, make the study applicable to our patient.

Resolution of Clinical Scenario

Jasinski et al. [12] reported an unbiased assessment of risk in their cohort. The authors provided long-term freedom from reoperation estimates and risk factors for poor outcomes in patients resembling our patient. There were some limitations related to study design and loss to follow-up, but we can apply the results to our patient and answer his question about long-term prognosis. Since the study included only a single surgeons' outcomes, the long-term data reported should be compared cautiously with our surgical skills and capability.

So, we can tell our patient that—given the nuances of bicuspid aortic valve repair and the complexity of this rare procedure—his risk of mortality at a mean follow-up of 5 years, with a long-term follow-up of 17 years is approximately 0.5%. As well, his freedom from reoperation during the same time period ranges from 92% to 80%. Based on the information provided to him, the patient opted to undergo a surgical aortic valve repair, with hopes of avoiding long-term complications associated with the use of prosthetic valves.

Appendix

Database used:

PubMed Advanced Search Filter

Key Words:

Bicuspid + Aortic Valve + Repair

Advanced Filters:

Human

Full Texts

Adults (18 years of age or older)

English

Time:

2021–2023 (within the last 3 years).

References

1. Moons Karel GM, Yvonne V, Grobbee Diederick E, Altman Douglas G. Prognosis and prognostic research: What, why, and how? *Br Med J*. 2009;338
2. Mack JW, Joffe S. Communicating about prognosis: Ethical responsibilities of pediatricians and parents. *Pediatrics*. 2014;133(Suppl 1):S24–30.
3. Randolph AG, Guyatt GH, Richardson WS. Prognosis in the intensive care unit: Finding accurate and useful estimates for counseling patients. *Crit Care Med*. 1998;26:767–72.
4. Italiano A. Prognostic or predictive? It's time to get back to definitions! *J Clin Oncol*. 2011;29:4718; author reply 4718–9
5. Barton S. Which clinical studies provide the best evidence? The best RCT still trumps the best observational study. *BMJ*. 2000;321:255–6.
6. Iorio ASF, Falavigna M. Use of GRADE for assessment of evidence about prognosis: Rating confidence in estimates of event rates in broad categories of patients. *Br Med J*. 2015;350:350.
7. Hansebout RR, Cornacchi SD, Haines T, Goldsmith CH. How to use an article about prognosis. *Can J Surg*. 2009;52:328–36.
8. Gamble JM. An introduction to the fundamentals of cohort and case-control studies. *Can J Hosp Pharm*. 2014;67:366–72.
9. Berger ML, Dreyer N, Anderson F, Towse A, Sedrakyan A, Normand SL. Prospective observational studies to assess comparative effectiveness: The ISPOR good research practices task force report. *Value Health*. 2012;15:217–30.
10. Thiese MS. Observational and interventional study design types; an overview. *Biochem Med (Zagreb)*. 2014;24:199–210.
11. Jasinski MJ, Kosiorowska K, Gocol R, Jasinski J, Nowicki R, Bielicki G, Berezowski M, Przybylski R, Obremska M, Lukaszewski M, Larysz A, Kansy A, Deja MA. Bicuspid aortic valve repair: Outcomes after 17 years of experience. *Eur J Cardiothorac Surg*. 2021;60:1053–61.
12. Mihailovic A, Bell CM, Urbach DR. Users' guide to the surgical literature. Case-control studies in surgical journals. *Can J Surg*. 2005;48:148–51.
13. Sleder A, Tackett S, Cerasale M, Mittal C, Isseh I, Radjef R, Taylor A, Farha R, Lupak O, Larkin D, Lamerato L, Divine G, Wisdom K, Baker-Genaw K, O'Neill W. Socioeconomic and racial disparities: A case-control study of patients receiving transcatheter aortic valve replacement for severe aortic stenosis. *J Racial Ethn Health Disparities*. 2017;4:1189–94.
14. Song JW, Chung KC. Observational studies: Cohort and case-control studies. *Plast Reconstr Surg*. 2010;126:2234–42.
15. Banerjee A, Chaudhury S. Statistics without tears: Populations and samples. *Ind Psychiatry J*. 2010;19:60–5.
16. Randolph AG, Guyatt G. Prognosis. In: Guyatt G, Rennie D, Meade MO, Cook DJ, editors. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. New York: McGraw-Hill; 2018.
17. Farrokhyar F, Bajammal S, Kahnemouli K, Bhandari M. Practical tips for surgical research. Ensuring balanced groups in surgical trials. *Can J Surg*. 2010;53:418–23.
18. Christenfeld NJ, Sloan RP, Carroll D, Greenland S. Risk factors, confounding, and the illusion of statistical control. *Psychosom Med*. 2004;66:868–75.
19. Akl EA, Briel M, You JJ, Sun X, Johnston BC, Busse JW, Mulla S, Lamontagne F, Bassler D, Vera C, Alshurafa M, Katsios CM, Zhou Q, Cukierman-Yaffe T, Gangji A, Mills EJ, Walter SD, Cook DJ, Schünemann HJ, Altman DG, Guyatt GH. Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): Systematic review. *BMJ*. 2012;344:e2809.
20. Jaffer IH, Whitlock RP. A mechanical heart valve is the best choice. *Heart Asia*. 2016;8:62–4.
21. Carlson MD, Morrison RS. Study design, precision, and validity in observational studies. *J Palliat Med*. 2009;12:77–82.
22. Walton MK, Powers JH 3rd, Hobart J, Patrick D, Marquis P, Vamvakas S, Isaac M, Molsen E, Cano S, Burke LB. International Society for Pharmacoeconomics and Outcomes Research Task Force for Clinical Outcomes Assessment. Clinical outcome assessments: Conceptual foundation-report of the ISPOR clinical outcomes assessment - Emerging good practices for outcomes research task force. *Value Health*. 2015;18:741–52.
23. Russell SD, Saval MA, Robbins JL, Ellestad MH, Gottlieb SS, Handberg EM, Zhou Y, Chandler B, HF-ACTION Investigators. New York Heart Association functional class predicts exercise parameters in the current era. *Am Heart J*. 2009;158:S24–30.
24. Rich JT, Neely JG, Paniello RC, Voelker CC, Nussenbaum B, Wang EW. A practical guide to understanding Kaplan-Meier curves. *Otolaryngol Head Neck Surg*. 2010;143:331–6.
25. du Prel JB, Hommel G, Rohrig B, Blettner M. Confidence interval or p-value?: Part 4 of a series on evaluation of scientific publications. *Dtsch Arztebl Int*. 2009;106:335–9.
26. Messé SR, Acker MA, Kasner SE, Fanning M, Giovannetti T, Ratcliffe SJ, Bilello M, Szeto WY, Bavaria JE, Hargrove WC 3rd, Mohler ER 3rd, Floyd TF; Determining Neurologic Outcomes from Valve Operations (DeNOVO) Investigators. *Circulation*. 2014;129(22):2253–61. <https://doi.org/10.1161/circulationaha.113.005084>. Epub 2014 Apr 1.
27. Gawande A. Two hundred years of surgery. *N Engl J Med*. 2012;366:1716–23.
28. Maganti M, Rao V, Brister S, Ivanov J. Decreasing mortality for coronary artery bypass surgery in octogenarians. *Can J Cardiol*. 2009;25:e32–5.

Alison B. Suraci and Gloria M. Rockwell

Abbreviations

BMI	Body Mass Index
ROM	Range of motion
RCT	Randomized controlled trials
QALY	Quality-Adjusted Life Years

Clinical Scenario

A junior orthopedic surgery resident assesses a 68-year-old female patient who presents to the emergency room with dehiscence and drainage from her surgical incision. She complains of increased pain with weight bearing over the last 5 days. She had an uncomplicated total hip arthroplasty 3 weeks prior. The patient lives at home independently with the help of family nearby. She ambulated without assistance aids prior to her hip replacement procedure. Her medical comorbidities include hypertension, type 2 diabetes, dyslipidemia, and osteopenia. Her body mass index (BMI) is 31. She does not have a smoking history. Her clinical exam demonstrates an aligned

lower extremity with full range of motion (ROM) of the hip and with distal intact neurovascular status. Radiographs show a congruent articulation and well-positioned implants without evidence of periprosthetic fracture. She does not have signs of systemic infection. Routine investigations with bloodwork inclusive of CBC, CRP, and ESR are above thresholds for suspicion of acute prosthetic joint infection. A hip aspirate performed under image guidance has a cell count diagnostic of acute infection, synovial fluid cultures are pending. The junior resident discusses the case with the chief resident, anticipating that a surgical intervention would be required. The chief resident explains the difference between acute (<6 weeks from index procedure) and chronic (>6 weeks from index procedure) infection as well as the approach to treatment with options being: (1) single surgery consisting of open debridement with retention of the existing prosthetic components, (2) single surgery with one-stage exchange of components, the complete explant of existing prosthesis and implantation of new final components, and (3) planned discrete procedures with two-stage exchange of components, where temporary antibiotic-coated components are implanted and a second surgery for implantation of final components occurs 6–12 weeks later. All procedures involve administration of 6–8 weeks of IV antibiotics, plus/minus extended oral antibiotic administration thereafter. The junior resident asks the chief resident how to decide which tech-

A. B. Suraci
Department of Surgery, Division of Orthopaedic
Surgery, Banner Medical Group, Phoenix, AZ, USA
e-mail: alison.suraci@bannerhealth.com

G. M. Rockwell (✉)
Department of Surgery, Division of Plastic Surgery,
University of Ottawa, Ottawa, ON, Canada
e-mail: grockwell@toh.ca

nique is the best to perform for this patient, and in general. The chief resident discusses the approach to these cases is balanced between onset of infection (acute versus chronic), aggressiveness of therapy, and associated morbidity to the patient and their final functionality.

Literature Search

Using PubMed, the residents performed a literature search (see Chap. 4) using the following search terms: “Acute Infection Total Hip Arthroplasty” AND “Surgical Treatment” and “Systematic Review,” and a filter of years (2010–2024); this search yielded 103 articles. To make the task manageable, the chief resident suggested they narrow the search strategy and consider using an article with a decision analytic approach. They keep the same criteria for year but change their search terms to “Acute Infection Total Hip Arthroplasty” AND “Surgical Treatment” AND “Decision Analysis”; this search yields five articles ([Appendix](#)). The residents screen the titles and abstracts of the articles; one of the articles (Article number 5) has the term decision analysis in the title, and one article (Article number 4) has the term decision-making in the title. Of those two articles only the former, by Bedair et al. [1] represents a proper decision analysis using evidence from the literature, and the one likely to provide the answer for them.

Background on Decision Analysis in Surgery

Clinical decision-making involves weighing the risks and benefits associated with the various treatment alternatives available for the management of a clinical scenario. Clinicians are guided in their decisions by personal experience, and by their ability to critically appraise the evidence in the literature. Decision-making becomes more difficult when there are multiple variables that can influence the outcome, when evidence in the literature is conflicting, inadequate, poorly designed, or absent, and when clinical scenarios

differ from the conditions of available studies. Under these circumstances of uncertainty, a formal decision analysis can help in addressing the clinical problem [2–6].

Decision analysis is the translation of a complex clinical scenario and its component parts into a manageable model. Competing management strategies are quantitatively compared despite the presence of clinical uncertainty. All possible therapeutic options and potential outcomes associated with each strategy are identified and assigned a value [7].

Decision analysis starts with a clinical question that must have a defined population, intervention, and outcome. Decision models use clinical experience and the literature to design the decision tree with all possible treatment outcomes for a given strategy. The probability that a specific outcome will occur is established from the literature. The strength of the literature for these outcomes is best achieved through high-quality randomized control trials or systematic reviews combining studies in a meta-analysis to determine the “base case” probabilities with confidence intervals. Base case is the best estimate of the probabilities for each variable; in other words it is the value closest to the truth in the author’s estimation [6]. The base case is determined through combining all the known estimates of outcomes by using a meta-analysis [6]. A sensitivity analysis is used to determine the robustness of the base case and can be performed by repeating the analysis in a one-way or multi-way fashion to determine reasonable alternatives to the outcomes as determined by confidence intervals or other intervals of values for the various outcome states [2–4, 8].

Components of Decision Analysis

Decision Tree

Decision analysis is composed of several stages in the model design. The first is accumulating outcome possibilities, probabilities, and utilities that are integrated into a decision tree. Utilities are defined as a health state with 0 (zero) representing death, and 1 (one) representing perfect

health, with possible negative values. The decision tree incorporates the elements of the clinical problem in a schematic format that represents all the clinical options, becoming more complex if multiple decisions are required to reach the final outcome for each strategy. The details behind the design and construction of a decision tree can be found in various references [3, 4, 8–10].

A decision tree is a standardized graphical display that is oriented from left to right. The decision that is to be made is placed on the left and marked with a square node (the decision node) [9]; the clinical outcomes are displayed

to the right. The lines (branches) originating from the decision node represent the different clinical strategies that are being compared. The circle nodes represent chance nodes, or chance events; outcomes are illustrated as triangles or rectangles [9]. Numbers can be present on decision trees, when these numbers are on the lines, they represent probabilities; when they are near the triangles, they represent utilities [9]. Figure 23.1 illustrates a simple decision tree, and its parts. Figure 23.2 illustrates an example of a decision tree that was used in the decision analysis by Bedair et al. [1] for surgi-

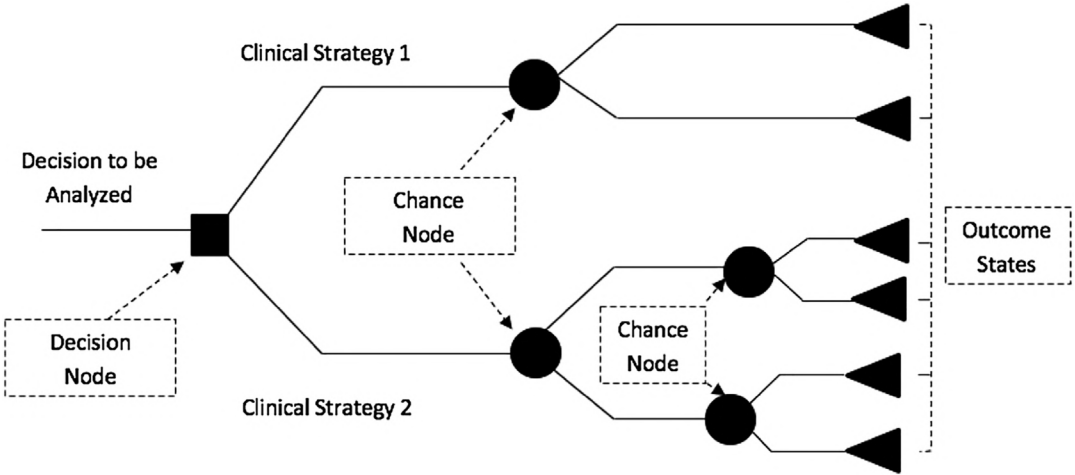


Fig. 23.1 Parts of a decision tree. (Created using information from Richardson and Detsky [9])

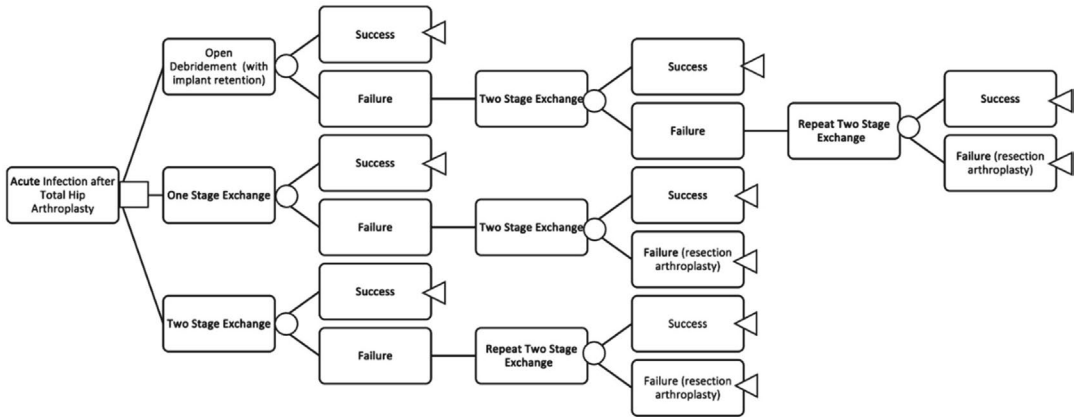


Fig. 23.2 The initial decision occurs at the root node (first decision point) at the left side of the tree where one of three treatment options is selected. Based on the chance (probability) of this treatment resulting in a specific outcome

(utility), the decision tree then describes sequentially (from left to right) possible further outcomes. The terminal node (right side of each pathway) describes the utility (quality of life) of the final result of that treatment path

cal treatment of early infection following total hip arthroplasty.

In the current example of surgical treatment options for acute infection following total hip arthroplasty; clinical outcomes were determined from the perspective of successful infection control. Failure was assumed to require subsequent surgery. A successful outcome was the control of infection at 2 years, with or without continued oral antibiotic suppression, determined by the need for subsequent surgical treatment. The model assumes the patient would undergo a maximum of two exchange procedures, and if infection eradication was still unsuccessful after the second exchange procedure, then would undergo resection arthroplasty in which no further implantation would occur. Figure 23.2 demonstrates a pictorial schematic of the decision tree for this algorithm. The three surgical choices included in the decision analysis are shown in Table 23.1.

The value of the given outcomes can be expressed in terms of utilities. The techniques to measure utilities for different health states can involve time trade-off, standard gamble, visual analog scales measured directly from experts and patients, or indirectly interpreted from similar clinical scenarios with previously published health utilities [4].

Disutility represents a transient health state that can downgrade quality of life temporarily and, for surgical interventions, could involve perioperative discomforts, inconvenience from hospitalization, and immobilization. Disutilities can be measured for short-term complications occurring in each health state. Disutilities are subtracted from the utilities of that health state to give an overall utility value for each specific outcome [4]. Permanent undesirable consequences decrease utility associated with a health state and can be converted to

Quality-Adjusted Life Years (QALY) for comparison to other health states, interventions, and even used in further economic evaluations such as cost-effectiveness analyses (see Chap. 24).

When the decision tree is folded back on itself (read right to left) the expected outcomes (utility, cost, QALY, etc.) can be calculated for each decision pathway by multiplying the probability of each branch in the tree with the final outcome for that pathway [4].

The Outcomes

All clinically relevant outcomes for the available management options must be identified, analyzed, and incorporated into the decision tree for the tree to be an accurate representation of a clinical scenario. Uncertainties (confidence intervals or other intervals) are identified during the literature systematic review. The sensitivity analysis of the tree is repeated by varying either one variable or multiple variables, throughout their plausible intervals, to determine the robustness of the decisions derived from the decision analysis [6]. Conclusions are not as strong if they change during the sensitivity analysis within the plausible interval of a variable. Extreme values can be used in the sensitivity analysis to debug the decision tree and help refine the design for obvious areas that give illogical results [4, 5].

Clinical Application

When applying a decision analysis clinicians must carefully consider the generalizability of the model to their own specific patient’s characteristics. The value of utilities a patient may apply to the various outcomes could vary and need to be considered when applying the decision tree.

Table 23.1 The four surgical treatment modalities

Treatment	Expected utilities
Open debridement (with retention of prosthesis)	0.86
One-stage exchange	0.82
Two-stage exchange	0.82
Resection arthroplasty	0.60

Adapted from Bedair et al. [1]

Users’ Guide: A Framework

Using the principles of decision analysis theory stated above and a previously devised methodological framework for use in interpreting the

literature, we can now critically appraise the Bedair et al. [1] article using the questions found in Box 23.1.

Box 23.1: Questions to Appraise Decision Analysis Literature

1. Are the Results Valid?
 - (a) Were all the important strategies and outcomes included?
 - (b) Was an explicit and sensible process used to identify, select, and combine the evidence into probabilities?
 - (c) Were the utilities obtained in an explicit and sensible way from the credible sources?
 - (d) Was the potential impact of any uncertainty in the evidence determined?
2. What Are the Results?
 - (a) In the baseline analysis, does one strategy result in a clinically important gain for patients? If not,
 - (b) Is the result a toss-up?
 - (c) Is the difference between the strategies clinically important?
 - (d) How strong is the evidence used in the analysis?
 - (e) Could the uncertainty in the evidence change the result?
3. Will the Results Help Me in Caring for My Patients?
 - (a) Do the probability estimates fit my patients' clinical features?
 - (b) Do the utilities reflect how my patients would value the outcomes of the decision?

Created using information from Refs. [6, 9, 10].

In their study, Bedair et al. [1] review the literature surrounding the surgical therapy for management of acute periprosthetic infection following a primary total hip arthroplasty. They expressed the advantages and disadvantages of each of the three most common surgical treatment options. Success of the procedure was defined as a period greater than 2 years without additional surgery for infection treatment, with or without adjunct long-term antibiotic suppression. Subsequent surgeries in the event of unsuccessful treatment of infection were protocolled in the model to assume as follows: (a) initial treatment with open irrigation and debridement, subsequent surgery with a two-stage exchange; and if this failed, repeat surgery with two-stage exchange; (b) initial treatment with one-stage exchange, subsequent surgery with two-stage exchange; (c) initial treatment with two-stage exchange, subsequent surgery with a repeat two-stage exchange. Complete failure of surgical treatment was defined as two exchange procedures, and ultimately the patient would then undergo resection arthroplasty without any reimplantation of components. This was based on a review of previous literature by Bedair et al. [1].

The authors used probabilities of obtaining a well-functioning hip arthroplasty with the described procedures based on previously published results [11–14]. The authors relied on the quality-of-life database compiled by the Institute for Clinical Research and Health Policy Studies for determination of utility values [15], and disutility values were deducted tolls from utility values. Initial surgery had one deduction from the final quality-adjusted life year and subsequent surgeries were further deducted tolls to account for disability from reoperation and associated lower health-related quality of life from multiple procedures (see Table 23.2).

Table 23.2 Probabilities, disabilities, and utilities by variable

Probability		Probability value
Successful treatment	Open debridement (with retention of prosthesis)	0.24 (0–0.99)
	One-stage exchange	0.93 (0–0.99)
	Two stage-exchange	0.92 (0–0.99)
Failure of initial treatment	Repeat two-stage exchange	0.60 (0–0.99)
	Two-stage exchange following failed debridement	0.92 (0–0.99)
	Repeat two-stage exchange following failed debridement	0.60 (0–0.99)
	Two-stage exchange following failed one-stage exchange	0.92 (0–0.99)
Utility		Utility value
Quality of life 1 = Perfect 0 = Death	Open debridement (with retention of prosthesis)	0.86
	One-stage exchange	0.82
	Two-stage exchange	0.82
	Resection arthroplasty	0.60
Disutility		Disutility value
Procedures	Open debridement (with retention of prosthesis)	–0.1
	One-stage exchange	–0.15
	Two-stage exchange	–0.2

Adapted from Bedair et al. [1]

Are the Results Valid?

Before applying the decision analysis to the patient in question, it is important to determine if the decision analytic model the authors applied is applicable to a real clinical scenario. This means examining the methodology of the model’s design, as to how the authors determined the outcomes, the methods used to combine the evidence, the sources and credibility of the utilities chosen, as well as the potential impact of uncertainty.

Were All the Important Strategies and Outcomes Included?

The first step in appraising a decision analysis is determining whether the authors have included all clinically relevant treatment strategies and outcomes. The decision analysis by Bedair et al. [1] reviews the literature of the three most common surgical interventions for acute prosthetic joint infection following total hip arthroplasty. Bedair et al. [1] tabulate the probability of suc-

cess for each procedure and associated quality of life following each independent procedure and subsequent procedures, mostly focusing on the probability of infection control and the toll of impact each procedure would have on expected quality of life. They provide a table of this determined from their review of the literature reported values. They clearly state the limitations of the available studies, and the inherent difficulty in comparison of studies by explaining the multiple variables stemming from differences in patient, surgery, and surgeon factors within each included study result in the treatment of this complex problem. The authors developed a decision model to fit a typical clinical scenario usually encountered in clinical practice, and inclusion of the algorithm for subsequent surgical treatment(s) based on current standards of care (Table 23.1). The failure of an initial surgery either an open debridement with retention of prosthesis, one-stage exchange, or two-stage exchange would result in the subsequent surgical treatment(s). Failure of treatment(s), defined as two revision/exchange procedures would result

in final treatment with resection arthroplasty as the endpoint.

How well the model fits with real clinical scenarios depends on reviewing the decision tree. Figure 23.3 is the complete decision tree with tabulated values. The treatment options are associated with their respective probability of success or failure. The terminal node represents the utility, or the health-related outcome measure of each treatment path (read left to right). The decision analysis is then reflected back on itself to demonstrate the value of each choice at decision points, and the optimal decision at the first node as the final result of optimizing the outcome being evaluated. The objective for the study was to determine the treatment option that results in the highest quality of life by balancing the surgical management that best controls infection while minimizing patient morbidity secondary to the impact of the procedure itself.

Was an Explicit and Sensible Process Used to Identify, Select, and Combine the Evidence into Probabilities?

The methods used to combine the evidence in a decision analysis are predictive of the clinical value of the final result. Providing the proper methodological steps are necessary to determine applicability to a given patient and also to mini-

mize the introduction of bias. Bedair et al. [1] state that they performed a literature review and that they were guided by these results to represent probabilities of success or failure of each treatment option [11–14]. They described the calculations for each outcome based on values reported in the literature for utility (quality of life) for uncomplicated and revision total hip arthroplasty procedures and their rationale for assigning a particular disutility value, as well as data limitations and assumptions used in the model. Unfortunately, their tabulated base case data for probabilities, utilities, and disutilities as well as the sensitivity analysis and threshold values, did not include confidence intervals or other intervals to more easily allow the reader to analyze the model and compare to their clinical case. Bedair et al. [1] report using the software program TreeAge Pro (Williamstown, MA, USA) [16] to model the clinical scenario described at the start of this chapter and construct the decision tree. Figure 23.2, showing the basic tree design moving from left to right, demonstrates that after each decision node for the three various surgical options, there was then a chance node of success or failure, with a further chance node of success or failure if subsequent surgical treatment was needed in effort for infection control. The values used by Bedair et al. [1] were critically analyzed within the materials and methods section of their

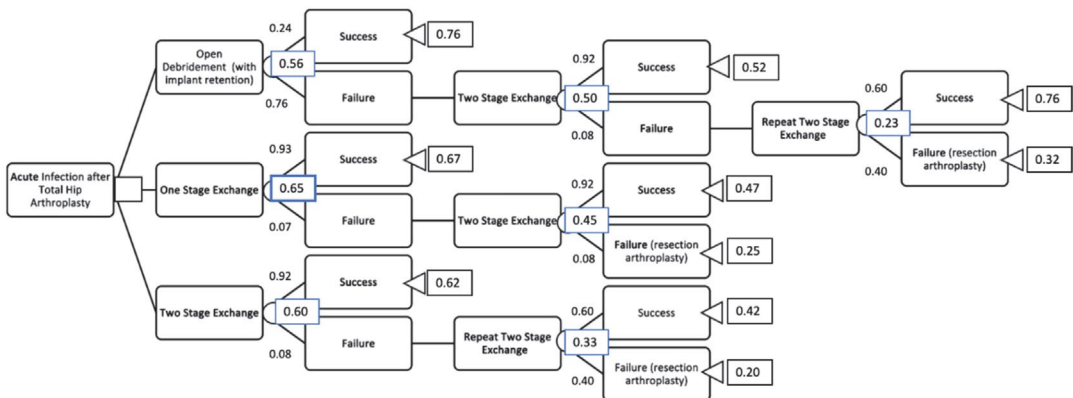


Fig. 23.3 The decision tree analysis shows the initial choice of one-stage exchange appears to result in the greatest expected benefit. Reflecting right to left along the branches of the tree, the utility at each branch is multi-

plied by its respective probability. The sum of these products along each branch that yields the highest number then will predict the pathway with the highest expected outcome. (Adapted from Bedair et al. [1])

paper. The authors chose base case success rates of 24% for open debridement with retention of prosthesis (14–74%), 93% (60–83%) for one-stage exchange, 92% for two-stage exchange (82–93%) and 60% for repeat two-stage exchange procedures, and 92% for two-stage exchange following either failed open debridement with retention or prosthesis or failed one-stage exchange [1]. Their model base case is biased toward a better outcome with exchange procedures, by choosing a lower initial base case success rate of 24% within an interval of 14–74% from their literature review for open debridement with retention of prosthesis and higher rates for exchange procedures, 92% varying from 60% to 83% for one-stage exchange and 93% varying from 82% to 93% for two-stage exchange [1]. Further analysis with a one-way sensitivity analysis was performed to determine threshold values (Table 23.3).

They assume the outcomes would not change with subsequent surgeries. The failure of previous surgery did not affect the success rate of the same but subsequent surgery and the disutility (toll) of subsequent surgery was assumed to be additive only. This is an assumption that may not necessarily be true.

Were the Utilities Obtained in an Explicit and Sensible Way from Credible Sources?

The authors used a simple decision analysis with two possible outcomes “success” (control of the infection at a period of greater than 2 years, without additional procedures for treating the infection) or “failure.” Utilities based on these outcomes were then included in this analysis to determine the optimal initial choice of surgical treatment. Several methods are available to measure utilities directly; different methods use different scales with zero equaling death and one equaling perfect health. Patient-specific utilities can be measured directly from the patient. Clinical policy utilities can be measured directly from large groups of patients with the disorder, from published studies of quality of life ratings by patients, and from large groups of people representing the general population.

In the decision analysis by Bedair et al. [1], utilities were adapted from the Clinical Research and Health Policy Studies, which is the largest reported database for utility (quality of life) after an uncomplicated total hip arthroplasty, recorded as being 0.86 [17]. The utility of resection arthroplasty was estimated to be 0.60 [18]. They assumed the utility for one-stage or two-stage procedure would be 0.82, although they do not explicitly state the derivation of this value. The impact of subsequent procedure to treat the infection was -0.1 for open debridement, -0.15 for one-stage exchange, and -0.2 for two-stage exchange procedures. These were assigned based on the authors evaluation of graded negative effect each procedure might have on the ultimate outcome. An example of the calculation is as follows: the utility (quality of life) after open debridement with retention of the prosthesis would be 0.86, assuming that quality-of-life following this procedure would be equal to that of an uncomplicated total hip arthroplasty with the toll applied of -0.1 to account for the impact of undergoing the procedure and therefore final utility value would be 0.76. This assigned graded toll (disutility) values associated with the individual procedures or subsequent surgery in the scenario of failure may not represent those of either patients or the general public.

Was the Potential Impact of Any Uncertainty in the Evidence Determined?

Uncertainty in the evidence, and poor model design can lead to miscalculation of both the utilities assigned to outcomes and the probabilities of each health state. As in any critical appraisal of the literature, the highest quality of evidence comes from well-designed, adequately powered randomized controlled trials (RCTs), and meta-analysis derived from such RCTs. Authors should acknowledge the quality of the literature in their review. Much of the uncertainty in decision-making arises from a lack of valid evidence in the literature. Even when present, published evidence is often imprecise with wide intervals of estimates. In their study, Bedair et al. [1] report some of the limitations in the quality of the studies in their review, with wide intervals of success and

Table 23.3 Results of one-way sensitivity analysis comparing surgical treatment pathways

Procedure		Sensitivity analysis (Threshold at which treatment dominates)
Probability of success	Open debridement with retention of prosthesis	>60%
	One-stage exchange	>66%
	Two-stage exchange	NT
	Repeat two-stage exchange	NT
	Two-stage exchange after failed debridement	
	Repeat two-stage exchange after failed debridement	
	Two-stage exchange after failed one-stage exchange	

Adapted from Bedair et al. [1]
NT no threshold

failure. Sensitivity analysis is the systematic exploration of uncertainty in the data to see what effect varying estimates or probabilities for risks, benefits, and utilities have on the expected outcome and change of a clinical strategy. The less valid the methods or less precise the estimates from the literature, the wider the intervals that must be included in the sensitivity analysis to determine when a threshold level is reached. The threshold value is reached when changing one parameter (varied over the range included) would alter the outcome of the analysis. Meaning that there exists a high or a low value of a specific parameter that would lead to a change of the optimal pathway (best choice) through the decision tree, and this is determined to be the threshold value for that treatment choice/pathway. Threshold values can help guide the decisions on which treatment option to choose based on the circumstances or parameters most closely matching the current situation.

The final outcome stated in the decision analysis by Bedair et al. [1] accounts for the probability of success and estimate of surgical procedure(s) on overall quality of life; a one-stage exchange would result in the optimal result by balancing aggressiveness of treatment with best infection control and minimization of morbidity incurred by the patient (Fig. 23.3). Bedair et al. [1] used the range of 0–99% in attempts to limit the bias from the wide variance in small studies used to report these values. In their sensitivity analysis, the success of a procedure was varied over this range and demonstrated that an open debridement with prosthesis retention would require a success rate of greater than 60% to shift the model to favor this as the optimal surgical procedure resulting in

maximal quality of life and the success of a one-stage exchange would have to decrease to less than 66% to favor open debridement with prosthesis retention or two-stage exchange or repeat procedures. There was no threshold value at which a two-stage exchange or repeat procedures would result in the greatest quality of life, for instance, even if the success rate was 99% (Table 23.3).

What Are the Results?

*In the baseline analysis does one strategy result in a clinically important gain for the patient?
If not is the result a toss-up?*

The choice between various outcome strategies is based on the outcome with the highest value of expected utility. One calculates this by folding back the decision tree probabilities from right to left for each outcome. The decision analyst chooses the scale on which to measure the expected utilities to fit the clinical problem: Quality of life, reduction of mortality, or other gains in the remaining life expectancy. Quality of life and quantity of life can both be combined in calculating the QALYs or healthy year equivalents. The authors chose to present their outcomes in a table of expected utilities for each procedure (Fig. 23.2).

Is the Difference Between the Strategies Clinically Important?

A difference means that any given patient may gain more or less than others and this is where using a policy decision analysis loses the specific-

ity for an individual patient. In the study by Bedair et al. [1], they display expected utilities without considering additional patient factors and consequential impact. The existing literature typically involves small sample sizes, and patient characteristics cannot be grouped to make more specific conclusions. For example, young and healthy patients may expect a better outcome because of their ability to participate in more advanced rehabilitation; however, time required for treatment may result in loss of income and/or social relationships. This scenario can be contrasted with a more chronically ill patient that might not reach the same capacity for rehabilitation but the financial/personal impact of length of treatment may not be as serious. Clinical importance remains a challenge with regard to patients and determining if infection control or well-functioning hip is more important, and so it is understood this balance may have shifted weight depending on the patient and their personal definition of clinical importance.

How Strong Is the Evidence in the Analysis?

The strength of evidence used in creating a decision analysis depends on the quality of the studies from which the probabilities and utilities were estimated. Ideally, every probability at each node in the decision tree is supported by precise estimates from primary RCT data or meta-analysis of high methodological quality. Good decision analysis can still be performed with some imprecise or ambiguous data as long as most of the data are good and the analysts explain any limitations and plan for sensitivity analysis accordingly. The weaker the evidence is used in the analysis the weaker the overall inference one can make from the results [3].

The strength of the evidence cited by Bedair et al. [1] is marginal; this is reflected by the absence of randomized or controlled studies used to construct their decision analysis. Deterministic data uses indirect information and requires a wider sensitivity analysis to test the robustness of conclusions. High quality directly measured data from RCTs or meta-analysis of RCTs provide the strongest evidence on which clinical decisions can be based. However, in such circumstances,

where strong data are readily available, the use of decision analytic modeling may not be necessary, unless residual uncertainty exists. The lack of high quality data in peer-reviewed literature in regard to outcome of success and/or quality of life following these procedures for the treatment of acute prosthetic joint infection following total hip arthroplasty does diminish the effectiveness of the reported results in this study; however this can serve as a basic foundation or framework for developing future studies that can more robustly test this hypothesis.

Could the Uncertainty in the Evidence Change the Results?

After interpreting the results in the base case analysis, it is critical to test how vulnerable the results are to the variation in probabilities and key outcome measures. If minor changes in the probability or effectiveness measures affect the outcome in a way that would be clinically important to the patient in question, then the decision model may have errors in design and should be reconsidered. A sensitivity analysis is considered robust if changing the variables throughout its possible intervals do not change the strategy or conclusion.

In the Bedair et al. [1] article, the authors explored the strength of their model with a wide sensitivity analysis. The robustness of their conclusions is demonstrated by the clinically unlikely probability that a two-stage exchange would ever provide the greatest quality of life in considering the options for treatment of an acute prosthetic infection following a total hip arthroplasty. First, by the results of the one-way sensitivity analysis varying the success of the procedure from 0% to 99% and demonstrating that even at the highest ranges for probability of success with this procedure, it still would not favor the greatest quality of life. Second, by the application of a two-way sensitivity analysis, whereby the authors compared varying success rates of open debridement with retention of prosthesis and one-stage exchange in an attempt to determine if these two most influential variables taken together could shift the model to favor a two-stage exchange. The results of the two-way sensi-

tivity analysis demonstrated that a two-stage exchange would only be favored if the success rates of open debridement with prosthesis retention and one-stage exchange approached values of 38% and 69% respectively. Successful treatment with an open debridement and prosthesis retention would afford the highest quality of life (0.76 versus 0.67) however, the limitations on its success and possibility of subsequent surgeries risk the eventual utility, and the sensitivity analysis demonstrates this by determining the success rate would need to be greater than 60% versus its assigned rate of 24%. This reaffirms their conclusion that a one-stage exchange will result in the greatest quality of life from initial decision analysis and that a two-stage exchange is unlikely to be the optimal treatment option from a quality of life standpoint where an open debridement with prosthesis retention would be unlikely to result in infection control.

Will the Results Help Me in Caring for My Patient?

Can the Results Be Generalized to This Patient?

The perspective, design, and assumptions used in creating a decision analysis become critical in determining the generalizability of the model, and in particular its relevance to the target population. The model must be based on clinical scenarios that are similar, predictive, and helpful to the actual reality of the clinical question at hand. Probabilities and utilities must reflect the values and risks present in the decision-makers' environment.

The applicability of a decision analysis for policy, group, or patient-specific interventions depends on the clinical characteristics of patients for whom the analysis is intended. The analysis should have a description of the patients' conditions to allow for generalizability. If the authors do not describe the samples used in the studies you can search inclusion and exclusion criteria in the original references to see how similar the characteristics are to your patient. If the analysis is designed for patients with differing characteristics than yours, you can assess the sensitivity

analysis for the variable used, and intervals, to see where your patient may fall in this current model.

In the decision analysis by Bedair et al. [1], the authors state their objective is to determine the treatment option that maximizes infection control balanced by minimizing the morbidity experienced by the patient. They outline the advantages and disadvantages of the three procedures as well as the assumptions used in this decision model. Patient characteristics (demographics, preoperative level of function, comorbidity index) and infection specifics (identification of organism, resistant organism, multi-organism) are not included in the course of this analysis and typically are featured in clinical decision-making. The strength of generalization in this paper highlights that despite differences in patient and infection qualities, the balance of infection control and quality of life is important to consider; and these two objectives in deciding on surgical procedure are in conflict irrespective of specific clinical scenarios.

Do the Probability Estimates Fit My Patient's Clinical Features?

Every decision analysis should contain pertinent positive and negative characteristics for the patient population to which it is applied. Reviewing the sensitivity analysis in the article will permit the gauging of similarities with the patient in question. In this article, the authors clearly state the target population is those with an acute prosthetic joint infection following a total hip arthroplasty (<6 weeks). The generalizability of this decision analysis is therefore limited to this group of patients. Highlighting this is not to be applied to patients with a chronic prosthetic joint infection following a total hip arthroplasty (>6 weeks), or patients with a prosthetic joint infection following a total knee arthroplasty.

Do the Utilities Reflect How My Patients Will Value the Decision Outcomes?

Quality of life, despite having universally accepted quantitative values, is strongly influenced by individual values and social context.

Surgeon comfort, experience, and discussion of potential risks for various interventions can all influence patients' choices. In this decision analysis, all failures of infection control were treated with subsequent surgery in the form of repeat exchange procedures.

This may not be representative of the reality in any given clinical scenario and may impact on decision nodes and thereby the validity of the model. If the decision analysis is designed for an individual patient, then the utilities should be measured directly from that patient. If utilities are based on a large group of patients or members of the public, then the utility values may include those of your patient, but the intervals of values must be very broad. One-way or two-way sensitivity analysis can further see if the values of your patient will affect the final decision.

In this example, utilities were derived from the largest available database reported by the Clinical Research and Health Policy Studies [15], specifically that of quality of life following an uncomplicated total hip arthroplasty procedure (0.86). The quality of life following a revision procedure was assigned value by the authors (0.82). The effect of subsequent surgery was graded and assigned by the authors (range -0.1 to -0.2). All but the first value is merit based from the literature. The derivation of toll is graded appropriately, taking into consideration the aggressiveness of the procedure. The quality of life following a revision procedure is questionable as its value is only slightly less than an uncomplicated total hip arthroplasty procedure.

This decision analysis has limited generalizability to any patient due to the indirect calculation of utilities, which are neither patient- nor procedure-specific. However, the broad sensitivity analysis demonstrated robustness of their conclusions, despite a bias toward exchange procedures.

Resolution of the Scenario

After working through the above steps, the residents had a much clearer understanding of the approach to discussion of surgical management

options for treatment of an acute prosthetic joint infection following total hip arthroplasty. Despite the absence of a well-powered randomized controlled trial comparing the three most common surgical treatment strategies, the decision analysis by Bedair et al. [1], demonstrated fairly robustly that one-stage exchange would result in optimizing the best result for infection control and a well-functioning hip. The residents felt well-informed in discussing the different strategies with this patient, helping the patient to provide informed consent for the procedure.

Appendix: Articles Located During Literature Search

1. Shohat N., Goswami K, Tan TL, Fillingham Y, Parvizi J. Increased Failure After Irrigation and Debridement for Acute Hematogenous Periprosthetic Joint Infection. *J Bone Joint Surg Am.* 2019;101(8):696–703. <https://doi.org/10.2106/JBJS.18.00381>.
2. Shohat N, Goswami K, Tan TL, Yayac M, Soriano A, Sousa R, Wouthuyzen-Bakker M, Parvizi J; for the European Study Group of Clinical Microbiology and Infectious Disease (ESCMID), for the Study Group of Implant Associated Infections (ESGIAI), for the Northern Infection Network of Joint Arthroplasty (NINJA). 2020 Frank Stinchfield Award: Identifying who will fail following irrigation and debridement for prosthetic joint infection. *Bone Joint J.* 2020;102:11–19. <https://doi.org/10.1302/0301-620X.102B7.BJJ-2019-1628.R1>.
3. Seyler TM. Traditional Laboratory Markers Hold Low Diagnostic Utility for Immunosuppressed Patients with Periprosthetic Joint Infections. *J Arthroplasty.* 2019;34(7):1441–1445. <https://doi.org/10.1016/j.arth.2019.03.013>.
4. Johnson NR, Rowe TM, Valenzeula MM, Scarola GT, Fehring TK. Do Pre-Reimplantation Erythrocyte Sedimentation Rate/C-Reactive Protein Cutoffs Guide Decision-Making in Prosthetic Joint Infection? Are We Flying Blind? *J*

Arthroplasty. 2022;37(2):347–352. <https://doi.org/10.1016/j.arth.2021.10.028>.

5. Bedair H, Ting N, Bozic K, Della Valle CJ, Sporer S. Treatment of Early Postoperative Infections after THA: A Decision Analysis. *Clinical Orthopaedics and Related Research*. 2011;469(12):3477–3485. <https://doi.org/10.1007/s11999-011-2119-0>.

References

1. Bedair H, Ting N, Bozic K, Della Valle CJ, Sporer S. Treatment of early postoperative infections after THA: a decision analysis. *Clin Orthop Relat Res*. 2011;469(12):3477–85. <https://doi.org/10.1007/s11999-011-2119-0>.
2. Detsky AS, Naglie G, Krahn MD, Naimark D, Redelmeier DA. Primer on medical decision analysis: part I—getting started. *Med Decis Mak*. 1997;17(2):123–5. <https://doi.org/10.1177/0272989X9701700202>.
3. Detsky AS, Naglie G, Krahn MD, Redelmeier DA, Naimark D. Primer on medical decision analysis: part 2—building a tree. *Med Decis Mak*. 1997;17(2):126–35. <https://doi.org/10.1177/0272989X9701700202>.
4. Naglie G, Krahn MD, Naimark D, Redelmeier DA, Detsky AS. Primer on medical decision analysis: part 3—estimating probabilities and utilities. *Med Decis Mak*. 1997;17(2):136–41. <https://doi.org/10.1177/0272989X9701700204>.
5. Naimark D, Krahn MD, Naglie G, Redelmeier DA, Detsky AS. Primer on medical decision analysis: part 5—working with Markov process. *Med Decis Mak*. 1997;17(2):152–9. <https://doi.org/10.1177/0272989X9701700205>.
6. Mastracci TM, Thoma A, Farrokhyar F, Tandan VR, Cina CS. User's guide to the surgical literature: how to use a decision analysis. *Can J Surg*. 2007;50(5):403–9.
7. Davis Sears E, Chung KE. Decision analysis in plastic surgery: a primer. *Am Soc Plast Surg*. 2010;126(4):1373–80.
8. Krahn MD, Naglie G, Naimark D, Redelmeier DA, Detsky AS. Primer on medical decision analysis: part 4—analyzing the model and interpreting the results. *Med Decis Mak*. 1997;17(2):142–51. <https://doi.org/10.1177/0272989X9701700204>.
9. Richardson WS, Detsky AS. User's guides to the medical literature. VII. How to use a clinical decision analysis. A. Are the results of the study valid? *JAMA*. 1995;273:1292–5.
10. Richardson WS, Detsky AS, for the Evidence-Based Medicine Working Group. User's guides to the medical literature. VII. How to use a clinical decision analysis B. What are the results and will they help me in caring for my patients? *JAMA*. 1995;273:1610–3.
11. Crockarell JR, Hanssen AD, Osmon DR, Morrey BF. Treatment of infection with debridement and retention of the components following hip arthroplasty. *J Bone Joint Surg Am*. 1998;80:1306–13. <https://doi.org/10.2106/00004623-199809000-00009>.
12. Tsukayama DT, Estrada R, Gustilo RB. Infection after total hip arthroplasty: a study of the treatment of one hundred and six infections. *J Bone Joint Surg Am*. 1996;78:512–23. <https://doi.org/10.2106/00004623-199604000-00005>.
13. Haddad FS, Muirhead-Allwood SK, Manktelow AR, Bacarese-Hamilton I. Two-stage un cemented revision hip arthroplasty for infection. *J Bone Joint Surg Br*. 2000;82:689–94. <https://doi.org/10.1302/0301-620x.82b5.9668>.
14. Pagnano MW, Trousdale RT, Hanssen AD. Outcome after reinfection following reimplantation hip arthroplasty. *Clin Orthop Relat Res*. 1997;338:192–204. <https://doi.org/10.1097/00003086-199705000-00026>.
15. Garvin KL, Hanssen AD. Infection after total hip arthroplasty: past, present, and future. *J Bone Joint Surg Am*. 1995;77:1576–88. <https://doi.org/10.2106/00004623-199510000-00015>.
16. TreeAge Pro 2021, R1. TreeAge Software, Williamstown; software available at <http://www.treeage.com>.
17. Rasanen P, Paavolainen P, Sintonen H, Koivisto AM, Blom M, Ryyanen OP, Roine RP. Effectiveness of hip or knee replacement surgery in terms of quality-adjusted life years and costs. *Acta Orthop*. 2007;78:108–15. <https://doi.org/10.1080/17453670610013501>.
18. Fisman DN, Reilly DT, Karchmer AW, Goldie SJ. Clinical effectiveness and cost-effectiveness of 2 management strategies for infected total hip arthroplasty in the elderly. *Clin Infect Dis*. 2001;32:419–30. <https://doi.org/10.1086/318502>.

Achilles Thoma, Feng Xie, Emily Dunn,
Jenny Santos, and Charles H. Goldsmith

Abbreviations

QoL	Quality of Life
HRQL	Health-Related Quality of Life
PRO	Patient Reported Outcome
RCT	Randomized Controlled Trial
MCID/MID	Minimal Clinical Important Difference/Minimal Important Difference
QALYs	Quality Adjusted Life Years
CA	Cost analysis
CEA	Cost-Effectiveness Analysis
CUA	Cost–Utility Analysis
CBA	Cost–Benefit Analysis
OECD	Organization for Economic Cooperation and Development
TRAM	Transverse Rectus Abdominis Myocutaneous
DIEP	Deep Inferior Epigastric Perforator

ZCQ	Zurich Claudication Questionnaire
ODI	Oswestry Disability Index
QBPDs	Quebec Back Pain Disability Scale
LSS	Lumbar Spinal Stenosis
IDD	Interspinous Distractor Device

Introduction

Surgeons of all specialties are regularly introduced to new surgical techniques or approaches to solve surgical problems. These innovative techniques are disseminated via conferences, workshops, or publications in specialty journals. The typical surgeon faces difficulty in deciding whether to adopt a new surgical innovation when conflicting opinions by experts are presented. A surgeon needs to consider the opportunity cost

A. Thoma (✉)
Division of Plastic Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: athoma@mcmaster.ca

F. Xie
Department of Health Research Methods, Evidence,
and Impact, McMaster University,
Hamilton, ON, Canada

Centre for Health Economics and Policy Analysis,
McMaster University, Hamilton, ON, Canada
e-mail: fengxie@mcmaster.ca

E. Dunn
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: dunne4@mcmaster.ca

J. Santos
Southwestern Public Health (SWPH),
St. Thomas, ON, Canada

C. H. Goldsmith
Department of Health Research Methods, Evidence
and Impact, McMaster University, Hamilton, ON,
Canada
e-mail: goldsmith@mcmaster.ca

when adopting a novel surgical intervention and abandoning the one they currently use. Opportunity cost is defined as “the value of the forgone benefits” because the resource is not available for its alternative use [1].

New innovations in surgery are often touted by their proponents as being cost-effective with the recommendation to adopt them in clinical practice. Unfortunately, the term cost-effective is often misused in surgical literature. For example, Ziolkowski et al. [2] found that most economic evaluations published in plastic surgery and touted to be cost-effectiveness studies were simply cost comparisons.

A surgical technique or surgical approach to be considered as “cost-effective” must integrate the costs and effectiveness [3, 4]. Economic analysis is a set of formal, quantitative methods used to compare alternative strategies with respect to their resource use and their expected outcomes [4]. Economic evaluation is a unique study design just as randomized controlled trials (RCTs) and case-control studies are. As most surgeons do not have training in health economics, one can understand how terminology such as “cost-effective,” is often misused in surgical literature. However, this misuse may have direct consequences if surgeons adopt new techniques or surgical approaches when they are not actually cost-effective; inefficient use of scarce health care resources is a direct consequence. Data from the Organization for Economic Cooperation and Development (OECD) data estimate that 20% of health expenditure worldwide is wasted, resulting in minimal—if any—improvement of health outcomes [5]. Although we do not have specific breakdown figures for surgery, we have no reason to believe that surgery is immune to wasteful practices.

This chapter has three objectives. First, it will introduce surgeons to the terminology used in economic evaluations and demystify this important study design. Secondly and most importantly, it will help the reader appraise and understand an article that purports to be an economic evaluation in surgery. Thirdly, it will stimulate surgeons to consider “piggy-backing” economic evaluations to their effectiveness

studies. We believe that an economic evaluation alongside a robust RCT that compares a novel surgical technique to a standard technique provides the best level of evidence to adopt or reject a novel surgical intervention. We will attempt to keep the mathematics to a minimum and make the chapter understandable and hopefully fun to read.

Types of Economic Evaluations

There are four types of health economic evaluations: Cost analysis (CA): this is a cost comparison study and usually not considered a complete economic evaluation, it is often called a partial economic evaluation, Cost-Effectiveness analysis (CEA), Cost-Utility Analysis (CUA), and Cost-Benefit Analysis (CBA) [4]. The main difference between these analyses is how the outcome or consequence(s) of the compared treatments are measured. The distinguishing features of these four economic evaluations are summarized in Table 24.1.

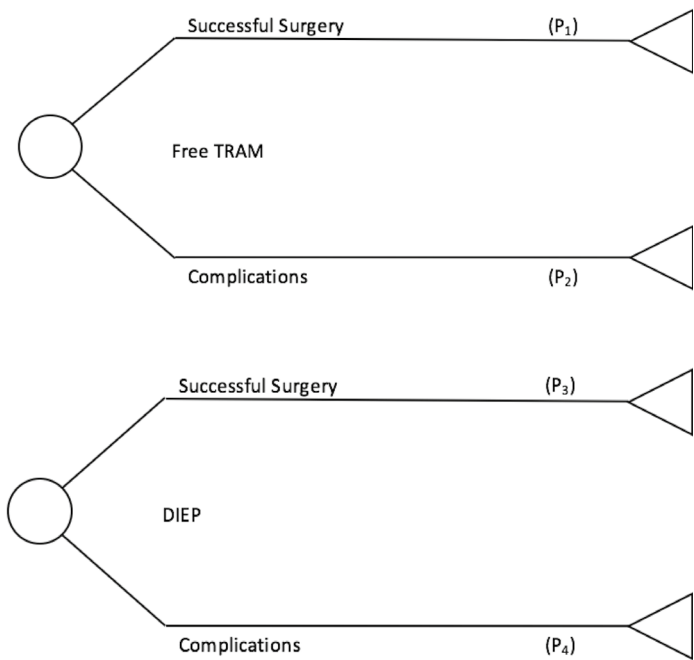
There are two main types of methodologies used in economic evaluations. The first type is the model-based evaluation and the second is the trial-based economic evaluation. In the model-based evaluation (also known as *deterministic analysis* or *probabilistic analysis*), a model is built, which in its simpler form is the decision tree explained in Chap. 23: *Decision Analysis and Surgery* of this book. Primary data are usually derived from the literature, e.g., by pooling the evidence (preferably through a systematic review of the literature). From these pooled data, one derives the probabilities of complications or positive health outcomes labeled as “health states.” These health states are then entered into a decision analysis tree, which in its most basic form will look like the one illustrated in Fig. 24.1a.

Clinical investigators then proceed to estimate the expected costs and expected benefits of the “health states” of the interventions under study by multiplying the costs and consequences (outcomes) by their probability of occurrence. The decision analysis tree model becomes more complex as more branches are added to the main

Table 24.1 Distinguishing features of the four types of economic evaluations

Type of analysis	Valuation of costs	Identification of outcomes (consequences)	Metric used in the analysis
Cost analysis (CA)	Monetary units (i.e., \$, £, €)	None	None. This analysis is only a comparison of costs
Cost-effectiveness analysis (CEA)	Monetary units (i.e., \$, £, €)	Common effect of interest. Common outcomes to the competing surgical interventions but with different probability of success (i.e., lives saved, successful hernia repairs, viable flaps, viable replants, sick days averted, hospital days averted)	\$ per natural unit (i.e., \$ per successful replant, \$ per life saved, \$ per hospital day averted)
Cost-utility analysis (CUA)	Monetary units (i.e., \$, £, €)	Single or multiple effects that are not necessarily common to both interventions. Outcomes are measured in health utilities that are used to calculate Quality Adjusted Life Years (QALYs)	\$, £, or € spent per QALY
Cost-benefit analysis (CBA)	Monetary units (i.e., \$, £, €)	Single or multiple effects not necessarily common to both surgical procedures and are calculated in \$, £, or €	Monetary units (i.e., \$, £, €)

Fig. 24.1a Decision Analytic Tree illustrating possible health state (pathway) probabilities for Free Transverse Rectus Abdominis Myocutaneous (TRAM) flap versus Deep Inferior Epigastric Perforator (DIEP) flap. * *P* probability of each pathway. (Figure from Thoma et al., [6])

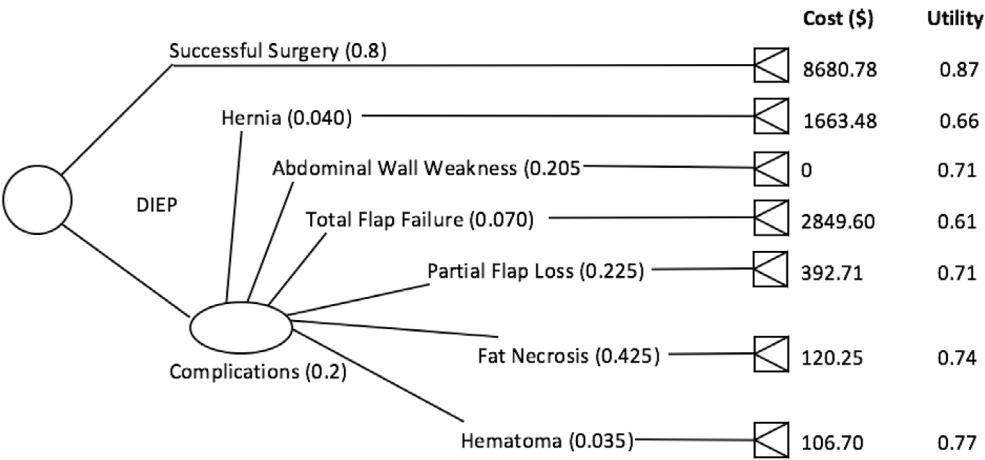


* *P*: probability of each pathway; figure from Thoma et al., (6)

pathways. This type of analysis is based on modeling. An example of a more complex model is shown in Fig. 24.1b, representing the cost and utility associated with various complications of a DIEP flap in postmastectomy breast reconstruction [6]. A similar analytic tree was also constructed for the other intervention for breast

reconstruction being studied; the free TRAM flap.

The second type, the trial-based economic evaluation (also known as *Stochastic Analysis*), derives data directly from trial participants. It incorporates sampling uncertainty that is inherent probabilistic by nature. Parallel to a prospective



*P: probability of each pathway; figure from Thoma et al., (6)

Fig. 24.1b DIEP flap in postmastectomy breast reconstruction. *P probability of each pathway. (Figure from Thoma et al., [6])

cohort study, and preferably, an RCT, the researchers capture not only clinical outcome data, but also direct and indirect costs related to the comparative interventions.

The adoption of a novel surgical intervention will be based on some threshold, which is arbitrarily set by clinical investigators or health economists based on consensus or in the case of CBA on showing a Net Social Benefit [4].

An important outcome that, unfortunately, is not commonly used in surgical comparative studies is Quality Adjusted Life Years (QALYs) [7]. This is an outcome that captures both quality of life and quantity of life in a single measure. It can be measured with preference-based instruments, such as the EuroQOL 5-Dimension (EQ-5D) and the Health Utilities Index [7] tools.

Formulas Used in Economic Evaluations that Inform Adoption of Novel Surgical Interventions

Cost Analysis

$$\text{Cost Analysis} = \text{Mean Cost}_{\text{novel surgical intervention}} - \text{Mean Cost}_{\text{comparative surgical intervention}}$$

(This is a partial economic evaluation. It assumes that the outcomes are the same, which may not be accurate unless one measures them and finds them to be so)

If the outcomes are similar, we adopt the novel intervention if it is found to be less costly.

Cost-Effectiveness Analysis

$$ICER = \frac{\Delta C}{\Delta E}$$

Here we calculate the Incremental Cost Effectiveness Ratio (ICER)

$$= \frac{(\text{Mean Cost}_{\text{novel surgical intervention}} - \text{Mean Cost}_{\text{comparative surgical intervention}})}{(\text{Mean Effectiveness}_{\text{novel surgical intervention}} - \text{Mean Effectiveness}_{\text{comparative surgical intervention}})}$$

(Although this is a full economic evaluation, the outcomes are measured differently and thus it cannot be used across disparate surgical interventions where outcomes are measured differently).

The main limitation of this type of analysis is that it is difficult to compare disparate surgical interventions. Imagine the hypothetical scenario where two specialties, plastic surgery and orthopedic surgery, present each to a third-party payer a CEA and ask that their novel surgeries be funded but there is funding for only one. If the ICER submitted by each specialty were as follows:

- ICER = \$40,000/successful breast reconstruction after mastectomy
- ICER = \$54,100/successful knee replacement

How should the third-party payer decide? This is akin to comparing apples and oranges. This dilemma is avoided by the CUA where a common outcome (QALYs) is used by both specialties and the decision is easier to make.

Cost-Utility Analysis

This is a full economic evaluation, recommended to be used to inform reimbursement policy making [3, 4]. Here we calculate the Incremental Cost Utility Ratio (ICUR).

$$ICUR = \frac{\Delta C}{\Delta U}$$

$$= \frac{\text{Mean Cost}_{\text{novel surgical intervention}} - \text{Mean Cost}_{\text{comparative surgical intervention}}}{\text{Mean QALY}_{\text{novel surgical intervention}} - \text{Mean QALY}_{\text{comparative intervention}}}$$

(If the ICUR < \$50,000/QALY, this is a strong indication that the novel intervention should be adopted [8]). As a general guidance, it is probably better to use a willingness-to-pay (WTP) threshold as the \$50,000 figure has no theoretical basis and it is mostly used in the USA.

Cost-Benefit Analysis

In this type of analysis, we attach a monetary value to the consequence of an intervention using a willingness-to-pay (WTP) approach.

$$NSB_t = \sum_{t=1}^n \frac{b_i(t) - c_i(t)}{(1+r)^{t-1}}$$

NSB = Net Social Benefit, t = year, b_i = benefits derived in year t, c_i = costs derived in year t, and r = discount rate, n = lifetime of study

(This type of economic evaluation is not used much in health care as it attaches a monetary value to the effectiveness of a health state, which health economists believe discriminates against the poor) [4].

Sensitivity Analysis

In clinical effectiveness studies, the results are reported by a point estimate such as a mean and a variation measure such as a standard deviation. In economic evaluations, one performs a sensitivity analysis to determine how robust the conclusions are [4]. After the base case analysis is performed based on the point estimate, the investigators redo the analysis based on one or two standard deviations around the point estimate of the costs or effectiveness or both, to see whether or not the conclusions of the study change. Probabilistic analyses are used to address parameter uncertainty in economic analyses [4, 9]. The conclusions are reported as robust if all calculations, main analysis, and sensitivity analyses favor the novel surgical intervention.

Now that you have mastered the important terminology and principles, we will help you understand how to appraise and understand a published economic evaluation report, by taking you through the following clinical scenario.

Clinical Scenario

At the weekly neurosurgical rounds, where the discussion was on Lumbar Spinal Stenosis (LSS), the newly hired surgeon, fresh from her Fellowship, asked why the X-Stop Interspinous Distractor Device (IDD) was not used in the patient under discussion. She claimed that in the center where she completed her fellowship, instead of the standard surgical decompression with Laminectomy, most spine neurosurgeons were using the IDD approach.

The chief of Neurosurgery, who is under pressure from the hospital administration to control the spending in his department, challenges her to provide the evidence that this new technology is indeed cost-effective so he can present it to the hospital administration.

Finding the Evidence

To identify the best evidence and inform her colleagues, the neurosurgeon begins by conducting a literature search, as described in Chap. 4: *Finding the Evidence Through Searching the Literature*, in this book and according to the “Users’ guide to the surgical literature: how to perform a high-quality literature search” [10].

The effectiveness of a surgical intervention can be found in a high quality RCT or a meta-analysis of number of RCTs. In this case, an RCT that compares Interspinous Distractor Devices (IDD) with Surgical decompression by Laminectomy in patients with Lumbar Spinal Stenosis, coupled with an economic evaluation, would provide the best evidence of cost-effectiveness. The surgeon follows the PICOT format, as described in Chap. 3: *Developing a Surgical Clinical Research Question*, for identification of important keywords used in the search process:

Population: Patients with lumbar spinal stenosis (LSS)

Intervention: Minimally invasive Interspinous Distractor Device (IDDs) insertion.

Comparison: Decompression by Standard Laminectomy

Outcome: Cost-effectiveness

Time horizon: At least a year follow-up

We performed a literature search by choosing the filtered database, COCHRANE reviews, and an unfiltered database, PubMed. With the PICOT format in mind, we used the search strategy: “Lumbar Spinal Stenosis (P) AND Interspinous Distractor Device (I) AND Laminectomy (C) AND cost-effectiveness analysis (O).”

The COCHRANE Search of these keywords identified one search result, Borg et al. [11]. The PubMed search identified two articles, including Borg et al. [11], and one additional article, which was a systematic review (see [Appendix](#)). We read

the abstract for Borg et al. [11] and we found it relevant to our purpose. This article includes both costs and effectiveness measured in QALYs and it is a full economic evaluation. It is an economic evaluation alongside an RCT that compared treatment by IDD or Laminectomy in 47 patients with LSS. Now, we proceed to appraise the economic evaluation to determine whether the results are valid and whether they are relevant to our practice.

Are the Results Valid?

Did the Analysis Provide a Full Economic Comparison of Health Care Strategies?

An economic analysis compares two or more health care interventions; in our case two surgical interventions to a surgical problem. Specifically, it compares the costs and the consequences (outcomes) of these interventions. The Borg et al. [11] article is a full economic evaluation as they compared the costs in pounds and the effectiveness of the interventions, IDD, and Laminectomy, in QALYs using the EQ-5D questionnaire administered at 6-, 12-, and 24-month time points. It seems this is a full economic evaluation, and we proceed to the next question.

Were Relevant Viewpoints Considered?

In an economic evaluation, there are a few possible perspectives (viewpoints). By this we mean who bears the costs associated with the use of the new surgical procedures. In appraising an economic evaluation specifically, we ask who is benefiting from this study? Is it the patient, the hospital, the third-party payer, or the society or others? There are occasions where a novel intervention may be found to be cost-effective from one viewpoint but not necessarily from another. For example, a hospital may save money by discharging patients earlier after a surgical intervention compared to an older program when the patients are hospitalized for a longer period but

from the patient's point of view, this may be more costly if the patient's spouse must take time off work to care for them during recovery at home.

In the Borg et al. [11] study, the authors were not explicit on the perspective taken in the study, however, in reading the article it becomes clear that this multicenter study took place in the UK and the costs were based on the National Health Service in England. The perspective thus taken was that of the NHS. This perspective is legitimate as the NHS covers the costs of the two procedures. In this study the investigators could have also examined the patient's perspective. The patient's perspective could have included any out-of-pocket costs to the patient such as physical therapy, medications, or assistive devices not covered by the NHS. A societal perspective could have included the NHS service costs, the patient borne costs, and productivity costs (indirect costs) such as time off employment. We wonder why they did not include the patient and societal perspective. On occasions the productivity costs can be greater than the third-payer costs. We wonder what this study would have shown had the authors considered these additional perspectives.

The perspective taken in an economic evaluation may depend on the question asked. The 1996 panel in cost-effectiveness in health strongly recommended that the societal perspective is the most important and should be considered, if possible [3]. In 2016, the US second Panel on Cost-Effectiveness in Health and Medicine published its update of the 1996 guideline [12]. The most important advance has been the idea of the reference case [13]. With regard to the most important perspective, the 2016 US Panel on Cost-effectiveness recommended that the CEA should be conducted from the perspective of the Public Payer [13]. Borg et al. [11] used the NHS perspective, which met the recommendations of the second Panel.

Were All Relevant Clinical Strategies Compared?

When conducting an economic evaluation, it is important that the investigators compare all relevant strategies for the condition under investiga-

tion. For example, the investigators should be comparing a novel procedure to a standard one and not one that is rarely used. In addition, investigators should be considering patients of different baseline risks. For example, if a general surgeon is performing a CEA on a novel hernia repair technique in the military population, this may not be generalizable to a retirement community population. If a CEA is performed comparing two approaches to hand surgery, one should consider both, the Workers Compensation Benefit (WCB) patient and the non-WCB patient population, as the WCB patients are considered high-risk patients.

Borg et al. [11], in their Discussion section of their article, mentioned the existence of alternative surgical approaches for LSS such as the advanced or minimally invasive techniques. They mentioned that in some centers these approaches are considered by some authors as the gold standard [14, 15]. They decided however to perform a comparative economic evaluation between standard laminectomy and IDD as in the three institutions where this study took place, the common procedure for LSS is the laminectomy approach.

Borg et al. [11] did not report subgroup analysis based on risk probably because the sample size of the study was small. If they had a large sample size, they could have performed a subgroup analysis by “operated lumbar levels,” e.g., 1 level, 2 levels, 3 levels. They could also have performed subgroup analysis by site of the surgery, e.g., L2-3 versus L5-S1 or by employment status, e.g., retired versus still working.

based, although both methods have their pros and cons. In trial-based, participant level data are extracted from an RCT, in which the investigators collect costs in parallel to the RCT. This type of economic evaluation provides high internal validity but at the expense of external validity as the participants in the study may not be typical of community patients. If multiple RCTs exist, one can then pool the results into a meta-analysis, thus increasing generalizability because the pooled estimate of the effectiveness is derived from a wider spectrum of participants. To improve generalizability, one may relax the inclusion criteria in the RCT, thus including patients that represent the whole population, thus making this a pragmatic RCT and pragmatic economic evaluation.

The clinical effectiveness in the Borg et al. [11] article was measured by several scales, which included the generic scales: EQ-5D, SF-36 and the Disease specific scales ZCQ, ODI, and QBPDS.

Borg et al. [11] also uses QALYs for performing a cost–utility analysis, the preferred type of economic evaluation by policymakers. To use QALYs, health utilities (such as HRQL weights) are required [4, 7, 16]. Borg et al. [11] used the EQ-5D to estimate health utilities. A health utility score is anchored at 0 (death) and 1 (perfect health). The other variable taken into account to measure QALYs is the duration of the corresponding health state. We believe that the investigators measured the effectiveness of the competing interventions appropriately.

Were Costs Measured Accurately?

How the cost of the interventions is reported depends on the perspective taken, as certain aspects such as costs of physiotherapy will not be relevant if comparing a hospital perspective, as these services are usually provided outside the hospital. Reporting health care resource use in natural units and unit costs separately allows for appraisal and replication by others.

As mentioned previously, Borg et al. [11] used the British NHS perspective. The hospitals under

Were the Costs and Outcomes Properly Measured and Valued?

Was Clinical Effectiveness Established?

As mentioned in the introduction, there are two methodologically distinct types of economic evaluations, the model-based method and the trial-based method. The preferred method is trial-

the NHS are reimbursed for procedures using national tariffs based on Healthcare Resource Group codes. As these tariffs for reimbursement often differ significantly from the true hospital costs, the authors decided to use the price for operating room (OR) time per minute and the price per day of admission to estimate the cost per patient.

Direct medical cost estimates of the competing interventions should include surgeon's fee, anesthesiology fees, OR costs (using micro-costing if possible), the IDD costs and physiotherapy costs (if these are performed in the hospital and covered by NHS). Borg et al. [11] reported the cost of running the OR in the three different centers that participated in the study. This was reported as £/min (USD/min). They also provided the cost of the IDD. The comparative costs can be found in table format in the Supplementary Table 3 of their article. Unfortunately, this table does not shed light as to whether the Theater (OR) costs also includes the anesthesia costs. Investigators should be transparent on such costs, so other investigators in other countries can replicate the study in their jurisdiction.

Were Data on Costs and Outcomes Appropriately Integrated?

Studies that claim to be full economic evaluations often compare direct medical costs only, which does not necessarily mean that whichever one is the least costly is the most cost effective [2, 17]. Another common mistake is to take a ratio of cost and effect of the novel intervention and compare it to that of the comparative intervention.

To determine whether a novel surgical treatment is cost-effective, one needs to calculate an ICUR. This integrates the costs and effectiveness of the competing interventions, telling us what the extra unit of benefit is for each extra unit of cost. This is precisely what the investigators in the Borg et al. [11] study did. They calculated the marginal cost per marginal unit of utility. This measure divides the difference in the mean cost of the novel and comparative treatment by the difference in the mean effectiveness, which in this case is QALY.

The investigators could have also reported an incremental cost effectiveness ratio (ICER), to look at the other outcome, SF36, ZCQ, ODI, and QBPDS. This equation is identical to the one used for ICUR (described above) other than the denominator:

$$\text{ICER} = \frac{\text{Differences in costs between IDD and Laminectomy}}{\text{Differences in health effects between IDD and Laminectomy}}$$

These measures are critical because they represent the treatment that has the greatest incremental cost per unit gained in either QALY or pain and mobility measures from the other scales used. In other words, the higher the value of ICER/ICUR, the greater the additional cost to achieve additional unit improvement of health (patient health). Acceptance or rejection of novel surgical technologies is based on this. It also depends on the patient and the circumstances at hand, such as the rarity of the intervention and ability of the health care system to support it. Ultimately, some type of threshold of acceptability will be decided upon by a consensus of experts.

In the Borg et al. [11] study, there was also an estimate of the total cost for each patient over

the 2-year follow-up, and they provided a discount rate of 2% per year. Therefore, it seems that costs and outcomes were appropriately integrated.

Was Appropriate Allowance Made for Uncertainties in the Analysis?

It is imperative to determine whether the ICER/ICUR values actually represent cost-effectiveness, as there are many values that can fall around the mean. This can be accomplished by recalculating the ICER/ICUR using both the best- and worst-case scenarios, referred to as a sensitivity analysis. If the conclusion on cost-effectiveness stays

the same, it can be decided with more confidence that the treatment is cost-effective. Borg et al. [11] performed several sensitivity analyses. These included (a) As Treated analysis (b) without the cost of device (c) without outliers, and (d) with cost of reoperations. We believe they satisfied this criterion.

Are Estimates of Costs and Outcomes Related to the Baseline Risk in the Treatment Population?

As patients who are considered “high risk” are generally more likely to benefit from a treatment than those who are considered “low risk,” it is imperative to divide the population to reflect these groups to determine if there is in fact a difference in cost and benefit between groups.

Borg et al. [11] included a table outlining the baseline characteristics of the population by the intervention group (Laminectomy Group versus X-Stop Group) and they were very similar between groups. If they had larger sample, they could have categorized the continuous variables (i.e., age and BMI). They could have also assessed by population risk such as “level operated,” “levels operated,” and employment sta-

tus. This would provide a better sense of the different baseline risks or benefits of one treatment over the other for older individuals or those considered “over-weight” or “obese”, those who have less severe disease (number of levels involved) (KL grade 2), and patients in workforce or not.

What Are the Results?

What Were the Incremental Costs and Outcomes Between the Two Strategies?

Borg et al. [11] found that both groups showed an overall improvement in mean QOL as measured by the EQ-5D at 6 months when compared with the preoperative baseline. At the 12- and 24-months’ time points, however, there was significant improvement in the laminectomy group only.

A summary of the incremental costs and outcomes are reported in their article. A more detailed description of these were reported as supplementary material in Tables 3–7. We have taken the information from these tables and present here the results in a clearer format (Tables 24.2, 24.3, and 24.4).

Table 24.2 Cost comparison of X-Stop vs. Laminectomy

	X-Stop (IDD)	Laminectomy	Difference
Total mean cost £ (range)	5148 (915–4842 ^a)	2711 (1356–4800)	
QALY	0.81	0.92	–0.11, <i>p</i> = 0.77, CI –0.349 to 0.078
ICER	ICER = (5148–2711)/(0.81–0.92) = –22,247.27.		

Adapted from Borg et al. [11]

^aExcluding device cost from range

Table 24.3 Intention to treat analysis

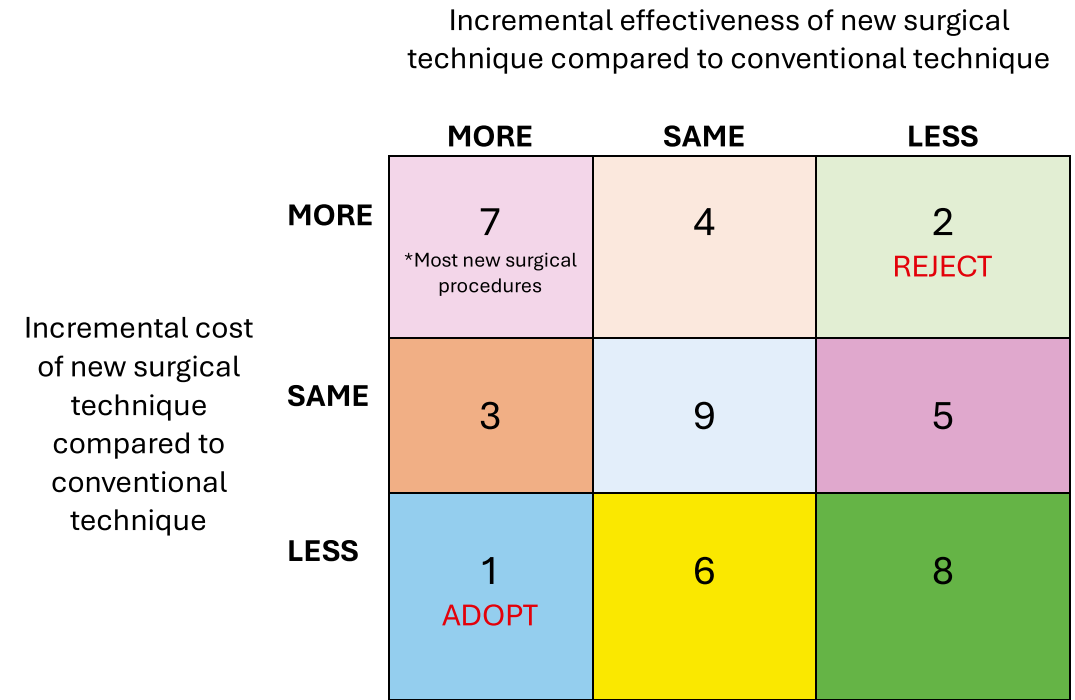
	X-Stop (IDD)	Laminectomy	ICER
Mean costs (£)	5148	2711	
QALYs	0.81	0.92	
			Laminectomy is dominant strategy. That is, it is less costly and provides more QALYs

Adapted from Borg et al. [11] supplementary Table 7

Table 24.4 Treated analysis

	X-Stop (IDD)	Laminectomy	ICER
Mean costs (£)	4941.98	2670.92	
QALYs	0.86	0.87	
			Laminectomy is dominant strategy. That is, it is less costly and provides more QALYs

Adapted from Borg et al. [11] supplementary Table 7



*Note Figure 2 was adapted, with permission, from Thoma et al. (18)

Fig. 24.2 Nine possible outcomes when comparing the new surgical technique and the conventional technique (the numbers within each cell are illustrative only). (*Note Fig. 24.2 was adapted, with permission, from Thoma et al. [18])

Do Incremental Costs and Outcomes Differ Among Subgroups?

Borg et al. [11] did not perform a subgroup analysis in high risk versus low risks patients and we cannot comment on this.

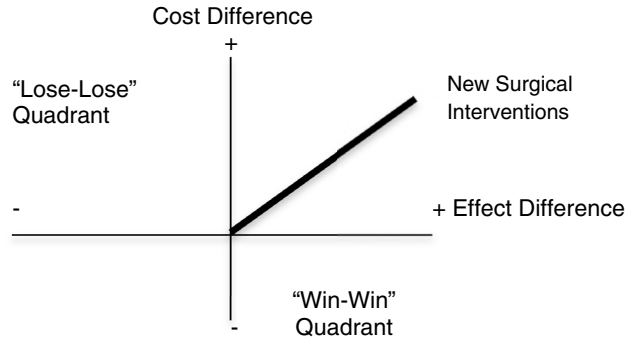
How Much Does Allowance for Uncertainty Change the Results?

Borg et al. [11] in their sensitivity analysis found that whether they analyzed their data as per intention to treat or as per treated analysis, the IDD was not an effective approach. Unfortunately,

they reported their result as $ICER = -£\ 22,145/QALY$ (see page 550 of their article). As reporting of economic evaluations with negative values is not intuitive, health economists prefer to use terms such as “dominant strategy,” “win-win” scenario, “lose-lose” scenario. In the Borg et al. [11] study, the authors should have reported that the IDD approach was a “lose-lose” strategy as surgery was more costly and less effective (Figs. 24.2 and 24.3).

When we compare a novel surgical intervention to a prevailing technology there are nine possibilities, which are illustrated in Fig. 24.2. The novel intervention may be more, same, or less effective than the prevailing

Fig. 24.3 Cost effectiveness plane.
(*Note: Fig. 24.3 was adapted, with permission, from Thoma et al. [7])



*Note: Figure 3 was adapted, with permission, from Thoma et al. (7)

intervention illustrated on the horizontal axis of Fig. 24.2. The vertical axis illustrates whether the novel intervention is more, same, or less costly than the prevailing intervention. If a novel intervention falls in *cell 1*, we adopt the new surgical technology, as it is more effective and less costly. Using the same reasoning, if it falls in *cell 2* we reject it, as it is less effective and more costly. Most novel interventions fall in *cell 7*, where new technologies are more effective but also more costly. Economic evaluations are needed in this category to find if the novel interventions are cost-effective. If a new technology falls in *cell 7* and the calculated ICER falls below an accepted WTP threshold, e.g., \$50,000/QALY, then the new technology is accepted. The Borg et al. [11] study found that the X-Stop interspinous Distractor Device costs more than Laminectomy. Furthermore, it was less effective. As such it would fall into *cell 2* in our Fig. 24.2, and we would reject it.

Another way of explaining these possibilities is the cost-effectiveness plane shown in Fig. 24.3, with the difference in effectiveness between the novel intervention and current treatment on the X-axis and the difference in cost on the Y-axis. If the novel intervention falls in the right lower quadrant, we have a win-win situation meaning it is more effective and less costly. Alternatively, if it falls in the left upper quadrant, we have a lose-lose situation meaning it is less effective and costs more. The slope of this line will determine the acceptability of the intervention. This

slope is a ratio, which is precisely what the ICER/ICUR represents.

Will the Results Help Me in Caring for My Patients?

Are the Treatment Benefits Worth the Harms and Costs?

When we compare two surgical interventions (novel versus standard), surgeons need to understand that there are nine possibilities, as explained in Fig. 24.2. If the novel intervention falls in *cell 1*, we accept the novel intervention, as it is more effective and less costly. If it falls in *cell 2*, we reject it, as it is less effective and more costly. We use a similar reasoning for the other scenarios. In most cases of modern surgery, however, most new innovations are more costly and effective. This is similar to a new intervention falling in the right upper quadrant of the cost-effectiveness plane (Fig. 24.3). It is under these circumstances that we need to perform a CEA or CUA. In the past, if an innovation had an ICUR of \$20,000/QALY the recommendation was given to accept it [19]. In recent years, this figure has increased \$50,000/QALY [19]. The above were proposed thresholds in the literature. A more official threshold is the one proposed by The National Institute for Health and Clinical Excellence (NICE), which suggests anywhere from \$27,000 to \$41,000 to accept an innovation [20].

Borg et al. [11] found that in their economic evaluations, which was “piggy-backed” to a pragmatic RCT, this new surgical approach, i.e., X-Stop was not cost-effective. Their data support this conclusion.

Could a Clinician’s Patient Expect Similar Health Outcomes?

In looking at Table 2 of the Borg et al. [11] article, and in particular the demographic characteristics of their patients pertaining to age, BMI, comorbidities, employment status, number of operated levels, and the level operated on, we believe that their patients are similar to ours and therefore have no reason to dispute their findings. If on the other hand our patients were much older and the majority of our patients had a different ratio of severity, we may be more skeptical of the findings shown here.

Can I Expect Similar Costs?

The costs of X-Stop and Laminectomy for LSS may differ in different jurisdictions (provinces, states, countries) and these should be considered seriously. If you believe that in your specific jurisdiction the costs are similar, then you should consider adopting the findings from this study. If on the other hand, the medical costs related to surgery are different in your geographic area, then you should recalculate the costs based on the health-related resource units provided by the authors of this paper. For example, you can calculate the cost per hospital admission in your hospital for the average patient by multiplying the number of hospital days found in this study. It is imperative for investigators not to just present the costs in an article but also state the resource units consumed for the two comparative interventions. From these data, we can recalculate the ICUR in our setting and decide for ourselves if the surgical option is cost-effective or not.

Resolution of the Scenario

Based on the baseline ICER, we see that the surgical approach fell into the lose-lose quadrant of the cost-effectiveness plane and therefore we are not prepared to accept this new approach, i.e., X-Stop for LSS for our patient. We also believe that the young neurosurgeon’s belief that this new innovation should be introduced to her hospital is wrong.

Final Thoughts and the CHEERS Statement

The EQUATOR network was established to enhance the quality and transparency of health research [21]. They provide guidelines or checklists for various types of research including randomized trials, systematic reviews, and economic evaluations.

Clinical investigators who perform economic evaluations are encouraged to report their studies by following the CHEERS statement [21]. The CHEERS statement is a guideline, specifically a checklist, that covers all methodological aspects of an economic evaluation (Title, introduction, methods, results, discussion, etc.). Although this guideline is not an appraisal instrument in itself, it does ensure that investigators cover all the elements of an economic evaluation. Familiarization with this guideline will ensure better quality study and eventual report.

Appendix

Cochrane Search Results ($n = 1$)

1. Borg A, Hill CS, Nurboja B, Critchley G, Choi D. A randomized controlled trial of the X-Stop interspinous distractor device versus laminectomy for lumbar spinal stenosis with 2-year quality-of-life and cost-effectiveness outcomes. *J Neurosurg Spine*. 2021;34(4):544–552.

PubMed Search Results (n = 2)

1. Ekhtor C, Griep D, Urbi A, Fiani B. Effectiveness of X-stop interspinous distractor device versus laminectomy for treatment of lumbar stenosis: A systematic review and meta-analysis. *Cureus*. 2023;15(4):e37535. <https://doi.org/10.7759/cureus.37535>.
2. Borg A, Hill CS, Nurboja B, Critchley G, Choi D. A randomized controlled trial of the X-Stop interspinous distractor device versus laminectomy for lumbar spinal stenosis with 2-year quality-of-life and cost-effectiveness outcomes. *J Neurosurg Spine*. 2021;34(4):544–552.

References

1. Chatterjee A, Payette MJ, Demas CP, Finlayson SRG. Opportunity cost: a systematic application to surgery. *Surgery*. 2009;146(1):18–22.
2. Ziolkowski NI, Voineskos SH, Ignacy TA, Thoma A. Systematic review of economic evaluations in plastic surgery. *Plast Reconstr Surg*. 2013;132(1):191–203.
3. Gold MR, Siegel JE, Russell LB, Weinstein MC, editors. *Cost-effectiveness in health and medicine*. New York: Oxford University Press; 1996.
4. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. *Methods for economic evaluation of health care programmes*. 3rd ed. Oxford University Press; 2005.
5. OECD. *Tackling wasteful spending on health*. [Internet]. Paris: OECH Publishing; 2018 [cited 2018 April]. Available from: <https://doi.org/10.1787/9789264266414-en>.
6. Thoma A, Veltri K, Khuthaila D, Rockwell G, Duku EK. Comparison of the deep inferior epigastric perforator (DIEP) and free transverse rectus abdominis myocutaneous (TRAM) flap in post-mastectomy reconstruction: a cost-effectiveness analysis. *Plast Reconstr Surg*. 2004;113:1650–61.
7. Thoma A, McKnight L. Quality adjusted life year (QALY) as a surgical outcome measure. A primer for plastic surgeons. *Plast Reconstr Surg*. 2010;125(4):1279–87.
8. Neumann PJ, Cohen JT, Weinstein MC. Updating cost-effectiveness—the curious resilience of the \$50,000-per-QALY threshold. *N Engl J Med*. 2014;371(9):796–7.
9. Thoma A, Kaur MN, Tsoi B, Ziolkowski N, Duku EK, Goldsmith CH. Cost-effectiveness analysis parallel to a randomized controlled trial comparing vertical scar reduction and inverted T-shaped reduction mammoplasty. *Plast Reconstr Surg*. 2014;134(6):1093–107.
10. Waltho DA, Kaur MN, Haynes RB, Farrokhyar F, Thoma A. Users' guide to the surgical literature: how to perform a high-quality literature search. *Can J Surg*. 2015;58:349–58.
11. Borg A, Hill CS, Nurboja B, Critchley G, Choi D. A randomized controlled trial of the X-Stop interspinous distractor device versus laminectomy for lumbar spinal stenosis with 2-year quality-of-life and cost-effectiveness outcomes. *J Neurosurg Spine*. 2021;34(4):544–52.
12. Neumann PJ, Sanders GD, Russell LB, Siegel JE, Ganiats TG, editors. *Cost-effectiveness in health and medicine*. 2nd ed. New York: Oxford University Press; 2016.
13. Krahn M, Bryan S, Lee K, Neumann PJ. Embracing the science of value in health. *CMAJ*. 2019;191(26):E733–6.
14. Lundberg J, Langevin JP. Lumbar microlaminectomy vs traditional laminectomy. *Fed Pract*. 2017;34(12):32–5.
15. Evaniew N, Khan M, Drew B, Kwok D, Bhandari M, Ghert M. Minimally invasive versus open surgery for cervical and lumbar discectomy: a systematic review and meta-analysis. *CMAJ Open*. 2014;2(4):E295–305.
16. Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. *Br Med Bull*. 2010;96(1):5–21.
17. Thoma A, Ignacy T, Ziolkowski N, Voineskos S. The performance and publication of cost-utility analyses in plastic surgery: making our specialty relevant. *Can J Plast Surg*. 2012;20(3):187–93.
18. Thoma A, Sprague S, Tandan V. Users guide to the surgical literature: how to use an article on economic analysis. *Can J Surg*. 2001;44:347–54.
19. Laupacis A, Feeny D, Detsky AS, Tugwell PX. How attractive does a new technology have to be to warrant adoption and utilization? Tentative guidelines for using clinical and economic evaluations. *CMAJ*. 1992;146(4):473–81.
20. McCabe C, Claxton K, Culyer AJ. The NICE cost-effectiveness threshold: what it is and what that means. *Pharmacoeconomics*. 2008;26(9):733–44.
21. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, Augustovski F, Briggs AH, Lindsay WR, Mauskopf J, Loder E. Consolidated health economic evaluation reporting standards (CHEERS) statement. *Eur J Health Econ*. 2013;14(3):367–72.



Studies Reporting Harm in Surgery

25

Lily J. Park, Marko Simunovic, Pablo E. Serrano,
Mohit Bhandari, and Robin McLeod

Dedicated to the memory of Dr. Robin McLeod—a shining light and leader for many years in Canadian surgery and in the development of Evidence-Based Surgery. This chapter borrows heavily from its previous iteration, authored by Dr. McLeod.

Abbreviations

ERAS	Enhanced Recovery After Surgery
NSAIDs	Non-Steroidal Anti-Inflammatory Drugs
DVT	Deep Vein Thrombosis
RCT	Randomized Controlled Trial
PICOT	Population, Intervention, Comparison, Outcome, Time Period
SUPREME	The Seoul National University Hospital Patients Research Environment
ASA	American Society of Anesthesiologists (ASA)
BMI	Body Mass Index
AJCC	American Joint Committee on Cancer

POD	Postoperative day
OR	Odds Ratio
CI	Confidence Interval
NNH	Number needed to harm
NNT	Number needed to treat
COX	Cyclooxygenase

Clinical Scenario

You are the Surgical Quality Leader at your hospital, you and your colleagues are interested in implementing an Enhanced Recovery after Surgery (ERAS) program. From your reading, ERAS protocols lead to improved outcomes, fewer complications, and decreased length of stay. You have organized several committees composed of individuals from all of the peri-operative groups that provide care to surgical patients. One of these

L. J. Park
Division of General Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada

Department of Health Research Methods, Evidence,
and Impact, McMaster University, Hamilton,
ON, Canada

e-mail: lily.park@medportal.ca

M. Simunovic
Division of General Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: simuno@mcmaster.ca

P. E. Serrano (✉)
Division of General Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence & Impact, McMaster University,
Hamilton, ON, Canada
e-mail: serrano@mcmaster.ca

M. Bhandari
Department of Surgery, Division of Orthopedic
Surgery, McMaster University,
Hamilton, ON, Canada
e-mail: bhandari@mcmaster.ca

R. McLeod
Department of Surgery and the Institute of Health
Policy, Management and Evaluation, University of
Toronto, Toronto, ON, Canada

groups is the Pain Management team, which is recommending multimodality pain medication to decrease the need for opioids. The multimodal analgesia package includes the prescription of non-steroidal anti-inflammatory drugs (NSAIDs). While you are supportive of a multimodality analgesia package, you recently heard a discussion at a conference on the use of NSAIDs in patients undergoing colorectal resections. While one of the colorectal surgeons supported the use of NSAIDs in this group of patients, another recommended being very cautious with the use of NSAIDs because of a possible increase in anastomotic leaks. You discuss this with your anesthesiologist colleague who feels quite strongly that NSAIDs should be part of the regimen. He points out that NSAIDs decrease the need for opioids, which, in turn decrease the likelihood of postoperative ileus. Furthermore, with recent concerns of opioid addiction, there is merit in decreasing their use. While you agree with your colleague, you point out that an anastomotic leak can be life threatening. After this discussion, you both agree that you need to review the evidence before making a recommendation.

Introduction

While surgery is performed to improve the health and quality of life of patients, there is always a risk that patients having surgery may develop adverse side effects. In some cases, the likelihood of a negative consequence is small compared to the benefits of surgery. In others, however, the risk of harm may be more significant because of the severity or frequency of adverse events. Surgeons must always be aware of both the benefits and harm of surgery and strive to ensure that the benefits of surgery outweigh any potential harm. In addition, they need to ensure that patients themselves understand the risks so they can participate in informed decision-making when it comes to their care. Just as surgeons must strive to use best evidence to assess the benefits of a surgical intervention, they should also use the best evidence to assess the risk of harm.

In this clinical scenario, both the benefits and risks of NSAIDs require assessment. While this might not be considered a risk of surgery per se, the

perioperative care of patients, which includes such interventions as preoperative antibiotics, deep vein thrombosis (DVT) prevention, and pain management, is an essential part of the overall care of surgical patients. In this example, the use of postoperative NSAIDs and the possible impact on the risk of an anastomotic leak, must be carefully evaluated by the surgeon and perioperative care team.

In line with the tenets of Evidence-Based Surgery, the best evidence for determining harm is required as much as it is needed for assessing benefit. Ideally, evidence from a randomized controlled trial (RCT) is the best evidence for making decisions about harm. However, RCTs are generally designed to evaluate the effectiveness of a treatment to improve patient outcomes and often only secondarily are they intended to measure complications or adverse events. This is due, in part, to the ethical concerns regarding randomization of patients to intentionally receive an intervention to determine its harm. In addition, some harmful effects may occur infrequently, and thus, the sample size would have to be very large to be able to make conclusions about the frequency of the harmful effects as the primary outcome. Furthermore, some side effects are not immediately apparent after treatment and therefore follow-up of patients participating in RCTs would have to be extended. For more information on RCTs see Chaps. 11, 12, 13, 14, and 15.

Instead of using RCTs to ascertain the risk of harm, cohort and case-control studies are most often used. Case-control studies are observational studies, which are often used in epidemiology. In these studies, individuals (cases) who are known to have the outcome of interest (harmful side effect) are matched to a cohort of patients (control) with similar baseline characteristics (diagnosis, type of surgery), but who do not have the outcome of interest. The advantage of these studies is that they require fewer resources, but the evidence is weaker (level 3 evidence) than that of a RCT (level 1 evidence) (see Chap. 5).

Case-control studies are often used to study rare diseases or as a preliminary study where little is known about the association between the risk factor and the disease of interest. One of the most notable use of a case-control study was the demonstration that smoking is associated with

Table 25.1 Description of the primary study designs

Characteristic	Study design		
	Randomized controlled trial	Cohort Study	Case-control study
Starting point	Intervention status	Exposure status	Event/outcome status
Group allocation	Randomization for intentional assignment to treatment groups; groups are balanced for known and unknown confounding factors	Groups are determined based on exposure status; groups may not be balanced	Groups are determined based on outcome status; groups may not be balanced
Outcome measures	Incidence of disease	Incidence of disease	Prevalence of disease
Measure of risk	Relative risk; odds ratio, risk difference	Relative risk; odds ratio; risk difference	Odds ratio
Temporal relationship between exposure and disease	Easier to establish	Easier to establish	Harder to establish
Strength	Bias controlled	Bias uncontrolled	Bias uncontrolled
Validity (if well designed)	Level I evidence	Level II evidence	Level III evidence

Adapted from Levine et al. [2]

lung cancer [1]. For more information on case-control studies see Chap. 18.

The other study design for determining whether a suspected risk factor leads to an adverse effect are cohort studies (level 2 evidence). Cohort studies are longitudinal studies in which patients are followed retrospectively or prospectively throughout a defined period of time. Unlike RCTs, the patients are not randomized to receive an intervention. Instead, patients are selected based on the exposure of interest, which has already occurred with no influence from the research team, to determine whether they will develop the outcome of interest. The limitation of cohort studies is that patients are not randomized and even if they are well matched in baseline characteristics between those exposed and unexposed, there may be some differences, which are not apparent or cannot be measured. The key strengths and limitations of the above study designs are summarized in Table 25.1. For more information on cohort studies see Chap. 17.

Finding the Evidence

To identify the best evidence, we performed a literature search. We first formulated our research question using the PICOT (population, intervention, comparison, outcome, time period) format as seen below.

- Population: Adult patients undergoing a colorectal resection with anastomosis
- Intervention: Any type of NSAIDs administered
- Comparison: No NSAIDs prescribed
- Outcome: Anastomotic leak
- Time horizon: Seven days following surgery

Using the PICOT format your research question would be: “*Does the use of any NSAIDs, as compared to no NSAIDs, increase the risk of anastomotic leaks in surgeries involving colorectal resection with anastomosis creation, seven days following surgery?*”

You perform a PubMed search based on the above-mentioned PICOT terms, you enter: “colorectal surgery” AND “non-steroidal anti-inflammatory drugs” AND “anastomotic leak” AND “postoperative” (for performing a literature search, see Chap. 4). You limit the search for studies published in the last 5 years, written in English, and involving adult patients (i.e., 19+ years) (Appendix). The original search yields four articles [3–6]. You do a title and abstract review of each of the four articles; two articles are the incorrect population, as they investigate NSAID use in esophageal and esophago-gastric surgery patients [4, 6]. There are two cohort studies, one of which primarily looked at reduction in the time to gastrointestinal recovery following colorectal surgery [3]. Since you are primarily interested in the potential harm of

NSAID, specifically as it relates to anastomotic leaks in colorectal surgery, you decide to review the last remaining study, entitled “Postoperative NSAIDs use and the risk of anastomotic leakage after restorative resection for colorectal cancer,” by Ju et al. [5].

Study Appraisal

To properly understand and present the information to your colleague, you have elected to use an article by Thoma et al. [7] to appraise your selected article. The questions used to appraise an article about harm in surgery are shown in Box 25.1.

Box 25.1: Questions Used to Appraise an Article Based on Harm

1. *Are the Results Valid?*
 - (a) Were patients similar in terms of prognostic factors that are known to be associated with the outcome (or was statistical adjustment necessary)?
 - (b) Were the circumstances and methods for detecting the outcome similar for patients and controls?
 - (c) Was the follow-up sufficiently complete?
2. *What Are the Results?*
 - (a) How strong is the association between exposure and outcome?
 - (b) How precise was the estimate of the risk?
3. *How Can I Apply the Results to My Patients or Clinical Practice?*
 - (a) Were the patients in the appraised study similar to the patients in my practice?
 - (b) Was follow-up sufficiently long?
 - (c) Is the exposure similar to what might occur in my patients?
 - (d) What is the magnitude of the risk?
 - (e) Are there any benefits that are known to be associated with exposure?

Critical Appraisal of the Article

A. Are the Results of the Study Valid?

Validity is a measure of the credibility of the research. Studies are valid if they actually measure what they set out to measure in a way such that the results can reasonably be applied to clinical practice. There are two types of validity: (1) internal validity, whether the study itself was conducted in a valid manner; and (2) external validity, whether the results of the study are generalizable. For example, if the goal of a study was to compare the effectiveness of two antibiotics in preventing postoperative wound infections it would be important that the patients in the two groups were similar in most, if not all, baseline and prognostic characteristics. If there were a larger proportion of patients in one group who had surgery for perforated diverticulitis, compared to the other group that included more patients who underwent simple hernia repairs without bowel resection, it would be difficult to conclude whether any differences in wound infection rates were from the antibiotics or due to the diverticulitis group having a higher risk of infection at baseline. We would conclude that this study lacked internal validity. On the other hand, if both groups were similar but included only patients who had elective operations for diverticular disease, we would be limited in our ability to generalize the results to patients having emergency surgery for perforated diverticulitis. This refers to external validity.

Were patients similar in terms of prognostic factors that are known to be associated with the outcome?

The Ju et al. [5] study reported data from The Seoul National University Hospital Patients Research Environment (SUPREME) system, a large national database. Information from this database and electronically registered medical records were used. For this study, only adult patients who had undergone an elective operation for colorectal cancer between January 2011 and December 2021 at a tertiary care center in Seoul, South Korea

were included. Exclusion criteria included patients without active cancer at the time of surgery, abdominoperineal resection, proctosigmoidectomy, and missing covariance data.

Demographic information, comorbidities, American Society of Anesthesiologists (ASA) class, body mass index (BMI), smoking status, neoadjuvant chemotherapy and radiotherapy, tumor stage, intraoperative blood loss, transfusion, type of procedure (laparoscopic or minimally invasive), and anastomotic leakage were collected in the SUPREME database using standardized definitions. The groups appear to be similar with most variables (age, BMI, neoadjuvant chemo- and radiotherapy, blood loss, transfusion) as well as the American Joint Committee on Cancer (AJCC) tumor stage, and resection type (rectal vs. colonic) (Table 25.2). However, there were differences in sex, smoking status, ASA class, preoperative serum albumin, duration of surgery, and whether the procedure was performed open or minimally invasive. Known risk factors for anastomotic leakage include sex (higher in males), smoking, obesity, alcohol abuse, and preoperative steroid use [8].

Were the circumstances and methods for determining exposure similar for patients and controls?

Both circumstances and methods appear to be similar for both groups. Generally, retrospective cohort studies have higher likelihood of bias compared to prospective cohort studies due to limited control in data collection. However, we can assume that the exposure (i.e., whether patients received NSAIDs) was well documented in the electronic medical records registry since all orders must be placed into this register. Considering this system, the circumstances and methods for determining exposure were likely similar in both the NSAIDs and no NSAIDs groups.

Was the follow-up sufficiently complete?

The standard follow-up to assess adverse events postoperatively in patients having colorectal surgery is 30 days. According to the literature, anastomotic leakage is most commonly detected 5–7 days following surgery [8]. In this study, there was complete follow-up in all patients at 30 days. You are therefore confident that the follow-up was an appropriate length, and sufficiently complete.

Was the correct temporal relationship demonstrated?

The authors report having classified the two groups based on whether NSAIDs were administered within 5 postoperative days

Table 25.2 Select patient characteristics

Demographic variables		No NSAID group (N = 7214)	NSAID group (N = 714)
Age	Median (interquartile range)	64 (56–72)	64 (55–73)
Sex	Female	2890 (40.1%)	320 (44.8%)
ASA status	III or higher	544 (7.5%)	79 (11.1%)
Body mass index kg/m ²	–	23.2 (21–25.4)	23.3 (21–25.7)
Current smoker	Yes	1027 (14.2%)	130 (18.2%)
Preoperative serum albumin, g/dL	Yes	4.1 (3.8–4.3)	4.1 (3.9–4.4)
Neoadjuvant chemotherapy	Yes	615 (8.5%)	72 (10.1%)
Neoadjuvant radiotherapy	Yes	14 (0.2%)	4 (0.6%)
American joint committee on cancer (AJCC) stage	Stage 3 or 4	3459 (47.9%)	341 (47.8%)
Resection	Colonic	2324 (32.2%)	255 (35.7%)
	Rectal	4890 (67.8%)	459 (64.3%)
Procedure	Open	4297 (59.6%)	360 (50.4%)
	Minimally invasive	2917 (40.4%)	354 (49.6%)

(POD) [5]. Since anastomotic leaks commonly occur at 5–10 days following surgery, limiting exposure of NSAIDs to within 5 post-operative days helps maintain temporality of any associations found in the results [9].

Was there a dose–response gradient?

Among the 714 patients that received NSAIDs within POD 5, 710 patients received intravenous ketorolac with a median and interquartile range dose of 30 (30–60) mg [5]. Two patients were coadministered intravenous ibuprofen, and one was co-administered oral aceclofenac. The remaining four patients received NSAIDs by way of intravenous ibuprofen only. Since most patients received 30 mg of intravenous ketorolac, the authors could not conduct a meaningful dose–response analysis, and so we cannot comment on a dose–response gradient.

What Are the Results?

How strong is the association between exposure and outcome?

The authors report that the anastomotic leak rate in the NSAID group was 1.7% whereas the anastomotic leak rate was 0.5% in the no NSAID group [5]. Thus, the increased absolute risk of an anastomotic leak with NSAID use was 1.2%.

Odds Ratios (OR) measure how strong an association there is between the exposure (NSAIDs) and outcome (anastomotic leak). An Odds Ratio of greater than 1 indicates that the risk of the outcome (anastomotic leak) is higher when exposed to the risk factor (NSAID) whereas an OR equal to one indicates that there is no association.

In this study, multivariate logistic regression analyses were performed including the following risk factors: NSAID use, age in years, sex, smoking status, concurrent diverting ostomy creation, duration of surgery, and intraoperative transfusion. The authors report that the adjusted OR of an increased risk of an anastomotic leak for patients who were administered NSAIDs was 3.41 (95% CI 1.76–6.60, $p < 0.001$) [5]. In other words, the risk of an anastomotic leak was increased 3.41-fold in patients taking NSAIDs postoperatively

[5]. Of note, the study re-demonstrated known risk factors for anastomotic leaks, as intraoperative transfusion and male sex were also associated with increased odds of experiencing an anastomotic leak, with statistical significance [8].

How precise is the estimate of the risk?

The precision of the risk is mainly affected by the sample size. The 95% CI will be wider with a small sample size and narrower in studies with a large sample size. In this analysis, the risk of anastomotic leak for patients taking NSAIDs was 3.41 and based on the 95% CI, we can be 95% certain that the risk lies somewhere between 1.76 and 6.60. In other words, based on this data, there is a high likelihood that taking NSAIDs is associated with an increased risk of anastomotic leak.

How Can I Apply the Result to My Patients or Clinical Practice.

Were the patients in the appraised study similar to the patients in my practice?

The characteristics of the patients in this study are representative of most patients you see for colorectal cancer surgery. However, there are some potential differences, although they likely do not affect the results. First, around 40% of the patients had laparoscopic surgery; this analysis included patients who had surgery between 2011 and 2021; the proportion of patients having surgery through a minimally invasive approach would, most likely, be higher now in most jurisdictions [10]. Second, this study was conducted through a National database in South Korea, so we must consider differences in the patient populations and management practices that can limit the external validity of the study findings. For instance, despite almost half the patients in both groups being categorized as AJCC stage 3 or 4, only a small percentage of patients received neoadjuvant chemotherapy (no NSAID group 8.5%, NSAID group 10.1%) and/or radiotherapy (no NSAID group 0.2%, NSAID group 0.6%). Especially with emerging evidence supporting the use of preoperative neoadjuvant therapy in advanced colon and rectal cancers, the patient population in this study

may differ from those in your usual practice [11, 12]. Furthermore, the anastomotic leak rates in this study population was low (0.5% and 1.7%) compared to previously reported rates of 2.5–12% [13–15]. Whether these outcomes are due to chance or effective institutional strategies to avoid leaks, the factors contributing to the overall low anastomotic rates are unknown. Nonetheless, the clinician should be aware of this and consider that as a result, effect estimates reported by the Ju et al. study (i.e., reported OR values), may have been overestimated, especially considering the ORs are larger than in previous studies [15].

Is the exposure similar to what might occur in my patient?

A variety of NSAIDs are available and may be prescribed to patients having elective colorectal surgery. This study suggests that further investigation is required to determine whether the risk of anastomotic leaks is lower in some NSAIDs and therefore may be safe for patients undergoing a colorectal resection. The exposure would therefore be similar for any patient that was prescribed NSAIDs; therefore, your patient would have the same risk of exposure as those patients in the Ju et al. [5] article.

What is the magnitude of the risk?

The Number Needed to Harm (NNH) is analogous to the Number Needed to Treat (NNT). However, while the NNT is a measure of how many patients need to be treated in order for one patient to benefit from the treatment, the NNH is a measure of how many people need to be treated (or exposed to a risk factor) for one person to have a particular adverse effect [16]. To determine the clinical importance of the results, it is worthwhile to calculate the NNH. If both are calculated, a low number for the NNT and a high number for the NNH would be the optimal finding [16]. For more information on magnitude and precision of treatment effects, please see Chap. 6.

The NNH is conceptually and mathematically simple for studies where there are distinct exposed

and unexposed groups. In this study, it can be calculated by using the raw data for anastomotic leaks in the three groups or use the percentages. Using the latter, the anastomotic leak rates were 1.7% in the NSAIDs group compared to 0.5% in the no NSAIDs group. Thus, to calculate the NNH, the anastomotic leak rate in the control group (0.5%) can be subtracted from the NSAIDs group leak rate: 1.7% minus 0.5% equals 1.2%. Then 100 divided by 1.2(%) is 83.3. So, it means that the NNH is 83.3 or in other words, for every 83 patients treated with diclofenac one additional patient would develop an anastomotic leak.

Are there any benefits known to be associated with exposure?

After reviewing the evidence regarding harm, we must weigh the potential benefits against the potential adverse events with the use of postoperative NSAID use. There are certainly benefits to using NSAIDs for pain in patients having colorectal surgery. While this study did not present data on pain control and length of postoperative ileus, others have shown that NSAIDs are effective in decreasing pain postoperatively and the need for opioids [16]. However, this study does show that the risk of an anastomotic leak is also increased, and it is well accepted that an anastomotic leak (especially those requiring reoperation) is one of the most common causes of postoperative mortality following a colorectal resection. Thus, this study provides evidence that there needs to be caution in prescribing NSAIDs postoperatively in patients who have had a colonic or rectal resection and anastomosis.

On the other hand, we also need more evidence to understand the benefits and risks of NSAIDs before our surgeon and anesthesiologist who are making the ERAS guidelines can make strong recommendations. For instance, this study showed that the risk was much lower in patients receiving NSAIDs (NNH 83), with all patients having received either ketorolac or ibuprofen, neither of which, are cyclooxygenase (COX)-2 selective NSAIDs. Theoretically, non-selective or COX-1 selective NSAIDs have a

higher risk of anastomotic leaks due to its impact on gastroduodenal toxicity. Therefore, although this study demonstrates increased risk of NSAID use on anastomotic leaks, the results are limited to nonselective NSAID use only. The authors did not do a temporal or dose-related evaluation, which might assist in making recommendations. Finally, we also know that the risk of anastomotic leaks is higher following rectal surgery than colonic resections so further evidence might change our recommendation in these cohorts of patients.

So what should the surgeon and anesthesiologist recommend? As in many decisions in surgery, there is a balance between the benefit and the risk. However, given the seriousness of an anastomotic leak, a cautious approach should be taken in the recommendation of NSAIDS.

Conclusion

Surgeons are frequently relied upon to provide information about surgical procedures to patients or, as in this scenario, hospital policies. In both instances, decisions should be made based on the best evidence, taking into consideration the benefits and risks of surgery as well as the patient's wishes or demands. Virtually every surgical procedure is associated with some amount of harm. If there is strong evidence regarding both the benefits and harm of the intervention, it is important that these outcomes are discussed with the patients. While the discussion should take into account the patients' desires, it is important that the information is provided in a way that can be understood by patients so they can make a decision.

Resolution of the Scenario

Although an RCT would provide stronger evidence, the methodology of this retrospective cohort study is strong, and they controlled for possible sources of bias between the two groups, which could render the results invalid. There was a significantly positive association between

NSAID use and anastomotic leaks following colorectal surgery. Indeed, the OR was 3.41 (95% CI 1.76–6.60) and almost all patients received intravenous ketorolac. Having this information, the surgeon and anesthesiologist decided to recommend that ketorolac should not be used for pain control in patients having a colorectal resection and anastomosis. However, they also noted that further studies are needed to understand which NSAIDs are not associated with anastomotic leaks and can be used with greater confidence in this cohort of patients.

Appendix: Literature Search Strategy

“colorectal surgery” AND “non-steroidal anti-inflammatory drugs” AND “anastomotic leak” AND “postoperative”

Limited to:

- Publication date: 2019-01-01 to 2024-01-01
- Article language: English
- Species: Human, Adult

References

1. Doll R, Hill AB. Mortality in relation to smoking: ten years' observations of British doctors. *BMJ*. 1964;1(5396):1460–7.
2. Levine M, Ioannidis J, Haines A, Guyatt G. Harm (observational studies). In: *Users' guide to the medical literature: a manual for evidence-based clinical practice*. 2nd ed. McGraw-Hill; 2008. p. 363–81.
3. Safety and efficacy of non-steroidal anti-inflammatory drugs to reduce ileus after colorectal surgery. Vol. 107. England; 2020.
4. Aiolfi A, Bonitta G, Campanelli G, Bona D. Impact of postoperative NSAIDs administration on anastomotic leak after esophago-gastric surgery: systematic review and meta-analysis. *Updat Surg*. 2023;75(4):817–24.
5. Ju JW, Lee HJ, Kim MJ, Ryoo SB, Kim WH, Jeong SY, Park KJ, Park JW. Postoperative NSAIDs use and the risk of anastomotic leakage after restorative resection for colorectal cancer. *Asian J Surg*. 2023;46(11):4749–54.

6. Corsini EM, Hofstetter WL. Ketorolac use and anastomotic leak in patients with esophageal cancer. *J Thorac Cardiovasc Surg.* 2021;161(2):448–54.
7. Thoma A, Kaur MN, Farrokhyar F, Waltho D, Levis C, Lovrics P, Goldsmith CH. Users' guide to the surgical literature: how to assess an article about harm in surgery. *Can J Surg.* 2016;59(5):351–7.
8. Daams F. Colorectal anastomotic leakage: aspects of prevention, detection and treatment. *World J Gastroenterol.* 2013;19(15):2293.
9. Li YW, Lian P, Huang B, Zheng HT, Wang MH, Gu WL, Li XX, Xu Y, Cai SJ. Very early colorectal anastomotic leakage within 5 post-operative days: a more severe subtype needs relaparotomy. *Sci Rep.* 2017;7:39936.
10. Pascual M. Laparoscopic colorectal surgery: current status and implementation of the latest technological innovations. *World J Gastroenterol.* 2016;22(2):704.
11. Cheong CK, Nistala KRY, Ng CH, Syn N, Chang HSY, Sundar R, Yang SY, Chong CS. Neoadjuvant therapy in locally advanced colon cancer: a meta-analysis and systematic review. *J Gastrointest Oncol.* 2020;11(5):847–57.
12. Kasi A, Abbasi S, Handa S, Al-Rajabi R, Saeed A, Baranda J, Sun W. Total Neoadjuvant therapy vs standard therapy in locally advanced rectal cancer: a systematic review and meta-analysis. *JAMA Netw Open.* 2020;3(12):e2030097.
13. Hyman N, Manchester TL, Osler T, Burns B, Cataldo PA. Anastomotic leaks after intestinal anastomosis. *Ann Surg.* 2007;245(2):254–8.
14. Reinke CE, Showalter S, Mahmoud NN, Kelz RR. Comparison of anastomotic leak rate after colorectal surgery using different databases. *Dis Colon Rectum.* 2013;56(5):638–44.
15. Lee L, Fiore JF. NSAIDs and anastomotic leak: what's the evidence? *Semin Colon Rectal Surg.* 2021;32(3):100833.
16. Andrade C. The numbers needed to treat and harm (NNT, NNH) statistics: what they tell us and what they do not. *J Clin Psychiatry.* 2015;76(03):e330–3.

Evaluating Surveys and Questionnaires in Surgical Research

26

Brian Chin and Christopher J. Coroneos

Abbreviations

BPI	Brachial plexus injury
COS	Core outcome sets
COMET	Core Outcome Measures in Effectiveness Trials
COSMIN	Consensus-based Standards for the selection of health Measurement Instruments

Introduction

Survey research is commonly used to gather information from a population of interest. When performed appropriately, surveys obtain important quantitative and qualitative data from a representative sample. In the surgical field, surveys are often used to assess the knowledge, perceptions, and attitudes of surgeons or patients to

capture a “snapshot” of clinical practice patterns at a point in time [1–3]. The results of such surveys can be used to plan future studies, for example: (1) a survey establishes a standard of care for the control group in a randomized controlled trial, or (2) a survey characterizes regional practice patterns to tailor an intervention in knowledge translation research. Although their use has been widely adopted, designing a survey requires careful planning to yield meaningful, unbiased results. The purpose of this chapter is to help surgeons understand the design and critical appraisal of surveys in the surgical literature.

Clinical Scenario

You are an American orthopedic hand surgeon with a practice focus on peripheral nerve injuries and brachial plexus reconstruction. While attending an international nerve surgery conference, you attend a presentation on the variation of surgical indications and techniques used to manage brachial plexus injuries (BPI). There is a lively discussion among the panelists and conference attendees representing many countries and surgical specialties including orthopedic surgery, plastic surgery, and neurosurgery. You wonder how your practice as an American orthopedic surgeon is similar or different to what other surgeons are doing in different specialties and countries.

B. Chin
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: chinh@mcmaster.ca

C. J. Coroneos (✉)
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
Department of Health Research Methods, Evidence,
and Impact, McMaster University, Hamilton,
ON, Canada
e-mail: coronec@mcmaster.ca

Literature Search

You visit [PubMed.gov](#) to perform a literature search with the terms: “brachial plexus injury” AND “survey” between 2018 and 2023. This yields 146 results in which two studies survey practice patterns for BPI. The first study [4] examines the role of therapy in conservative

management of pediatric BPI among expert therapists. The second paper by Lubelski et al. [5] is a survey comparing the surgical treatment of adult and pediatric BPI among peripheral nerve surgeons. Given its relevance, you decide to review the paper by Lubelski et al. [5] and evaluate its methodology (Table 26.1).

Table 26.1 Key characteristics of the survey by Lubelski et al. [5]

Title	Differences in the surgical treatment of adult and pediatric brachial plexus injuries among peripheral nerve surgeons
Study design	Cross-sectional survey
Survey development	Developed by authors in the Department of Neurosurgery, John Hopkins University School of Medicine
Pilot testing	Not reported
Clinical sensibility testing	Not reported
Reliability and validity testing	Not reported
Sample	Members of major peripheral nerve professional organizations
Method of administration	Email
No. of contacts	Once
Overall Response Rate	154/255 (60.4%)
Continent	North America: 86/154 (55.8%) Europe: 38/154 (24.7%) South America: 16/154 (10.4%) Asia: 11/154 (7.1%) Africa: 3/154 (2.0%)
Specialty	Neurosurgery: 51/151 (33.7%) Orthopedic Surgery: 49/151 (32.5%) Plastic Surgery: 49/151 (32.5%) General Surgery: 2/151 (1.3%)
Practice Type	Academic: 82/153 (53.6%) Private: 14/153 (9.1%) Hybrid: 57/153 (37.3%)
Years in practice	<10 years: 51/151 (33.8%) 10–19 years: 38/151 (25.2%) 20–29 years: 28/151 (18.5%) >30 years: 34/151 (22.5%)

Appraisal of a Surgical Survey

The appraisal of surgical surveys centers on evaluating the validity of the study, interpreting the results, and applying study findings clinically (Box 26.1) [6–8]. This framework will help readers determine whether the methodological development and execution of the survey are reliable to obtain accurate information, and whether the results are relevant to change practice.

Box 26.1: Framework for How to Evaluate a Surgical Survey [6–8]

I. Are the results valid?

Primary Guides

- (i) Is there a clear research question and objective?
- (ii) Was there an appropriate selection of the sampling frame?
- (iii) Was there appropriate development of the questionnaire (item generation, item reduction, formatting, and pretesting)?
- (iv) Was the administration of the questionnaire appropriate?

Secondary Guides

- (i) Was pilot testing performed?
- (ii) Was clinical sensitivity testing performed?
- (iii) Was reliability and validity testing performed?

II. What are the results?

- (i) Was there a sufficient response rate?
- (ii) Are appropriate statistical methods used?
- (iii) Is the reporting of the results transparent?
- (iv) Are the conclusions appropriate?

III. Will the results change practice?

- (i) Are the results generalizable to my practice?
- (ii) Will the conclusions from this survey change my practice/behavior?

I. Are the Results Valid?

Primary Guides

- (i) *Is there a clear research question and objective?*

A clear research objective is essential to survey design. Each questionnaire should be guided by a primary research question that highlights the topic of interest and the target respondents [8]. Posing a clinically relevant and interesting question is more likely to attract attention and elicit responses. The survey by Lubelski et al. [5] has clear objectives: to (1) identify the variability in the management of BPI, (2) determine factors associated with practice variability, and (3) provide an overview of the current state of BPI management.

- (ii) *Was there appropriate selection of the sampling frame?*

A survey should have a clearly defined target population. In an ideal scenario, a survey is administered to the entire *target population* to obtain the most complete and accurate data. This is often impractical due to the large size of the population or study constraints; a sample of the target population is often surveyed (Fig. 26.1) [9]. The *sampling frame* can be defined as the empirically measurable members of the target population from which the sample is drawn [7–10]. The *sampling element* refers to the respondents of the sampling frame whose responses are collected and analyzed [10]. There may be individuals of the target population that cannot be surveyed and thus fall outside the sampling frame. Thus, it is important to select a specific sampling frame that best captures the target population of interest.

The reader should consider if the sampling frame appropriately represents the target population, and how the individuals within the population are identified to be respondents. If a membership registry is used, readers must consider if individuals

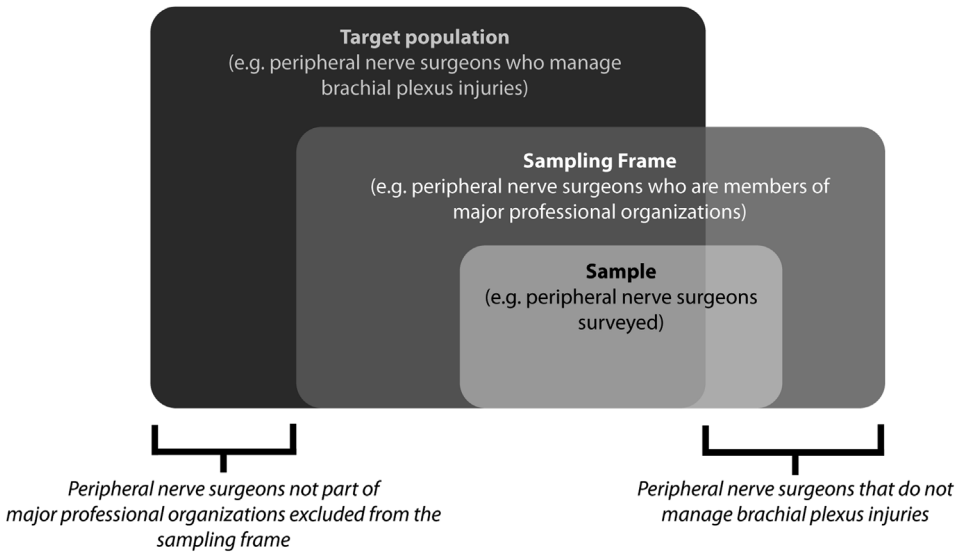


Fig. 26.1 Illustrative diagram of sampling strategy used by Lubelski et al. [5] for peripheral nerve surgeons. When selecting the sampling frame, consider individuals of the target population that may be poten-

tially excluded and how it will affect the generalizability of the survey. Also consider potential inclusion of individuals that are not part of the target population within the sampling frame

outside of the target population (e.g., fellowship trainees, surgeons no longer in practice) may be included. Random sample selection is based on probability designs, which include: simple random sampling, systematic random sampling, stratified random sampling, and cluster sampling [10, 11]. Deliberate nonprobability-based sampling is utilized to study populations that may be difficult to identify. Such designs include purposive sampling, quota sampling, chunk sampling, and snowball sampling [10]. Ultimately, the degree to which survey results can be generalized will depend on how similar the sample is to the target population of interest.

Lubelski et al. [5] chose peripheral nerve surgeons who manage BPI as the target population (Fig. 26.1). The sampling frame was the membership list of major peripheral nerve professional organizations, but the specific organizations are not stated. It is difficult to evaluate if the sampling frame is representative of the target population without knowing the

specific organizations. There may be a selection bias based on the national, continental, or international status of the organization. The authors also state that invited participants were selected after reviewing membership registry, which implies there was some form of selection process to determine which members managed BPI. Although selecting a membership registry of peripheral nerve surgeons likely represents the best possible sampling frame from a pragmatic perspective, further detail and clarity of the sampling frame would strengthen the authors' methodology.

(iii) *Was there appropriate development of the questionnaire?*

The framework for survey questionnaire development includes four major phases: (1) item generation, (2) item reduction, (3) formatting, and (4) pretesting [8].

Item generation involves conceptualizing the research question and identifying all potential ideas and concepts that could be included. During the process, individ-

ual items are organized into broader themes or domains relevant to the research question to ensure all key aspects of the clinical problem are addressed. Items can be generated through a systematic review, discussion with potential responders, expert panels, and focus groups. The Delphi technique is a well-recognized process of item generation and ranking to develop a consensus that can be utilized in questionnaire development [12]. Once item generation is exhausted (“sampling to redundancy”), individual items are grouped into larger domains wherein questions are developed.

Many potential questions within domains are pruned for redundancy and consolidated during item reduction, without eliminating entire domains or important constructs. It is critical to limit the number of questions in a survey, as lengthy surveys reduce response rates [13]. Item reduction is an iterative process achieved through ranking or rating of items with potential responders and content experts, or through external appraisal and statistical methods.

Survey questions should be formatted to be concise, easy to understand, and organized with a logical flow. The goal is to limit ambiguity in the question stem and response options. Question stems should be fewer than 20 words, with language appropriate for the lowest educational level in the cohort [12, 14]. Close-ended questions restricted to binary (e.g., yes or no) or limited responses (e.g., Likert scale: Never, Rarely, Sometimes, Frequently, Always) are ideal for analysis. In contrast, open-ended questions are challenging to aggregate and analyze [15]. Each question should focus on a single construct to avoid “double barrel” or compound questions, which involve more than one issue but only allow for one answer (e.g., How satisfied are you with your physician *and* medical care?) [7, 16]. When appropriate, indeterminate responses and “other”

options should be included to allow for uncertainty [8, 14]. Terminology considered judgmental or absolute (e.g., “always,” “never”) should be avoided as they may bias responses [17].

Lastly, pretesting should be conducted to ensure questions are interpreted appropriately and consistently by respondents. Pretesting is used to identify items that may be poorly worded, prone to misinterpretation, or difficult to answer [18, 19]. It is also an opportunity to check for unclear directions and assess the time to complete the questionnaire. Obtaining feedback from individuals who are similar to prospective respondents allows for questions to be further refined. Specific methods of formal questionnaire testing are detailed in the *Secondary Guides*.

The process of questionnaire development falls under the larger topic of measurement instruments, which is beyond the scope of this chapter. Factors to consider in questionnaire design are an appropriate selection of the variable to measure, and the most suitable instrument to measure it. The COMET (Core Outcome Measures in Effectiveness Trials) initiative has made a significant effort to develop core outcome sets (COS), which should be measured for research on a specific condition [20] (See Chap. 34). In kind, the COSMIN (CONsensus-based Standards for the selection of health Measurement INSTRuments) initiative provides the best-validated instrument to measure a specific outcome [21]. Consider utilizing these or similar resources to guide development of surveys and strengthen its methodology.

Lubelski et al. [5] used four clinical vignettes to elicit surgical management decisions of its participants. The authors state that the clinical vignettes are examples of upper and lower BPI patterns in adult and pediatric patients. The vignettes are complex patient presentations with varying injury patterns, time since injury, clinical recovery, and diagnostic test

results. Based on these vignettes, participants were asked to decide on operative management, timing of surgery, and what procedures they would perform. Despite the enormous complexity of the survey, review of the methods reported by Lubelski et al. [5] provides minimal detail on the process of survey development. There is no information regarding the process of item generation, reduction, or pre-testing. Individual questionnaire items could not be assessed since the survey was not available for review in the main text or supplementary documents, but the general format could be inferred based on study results. Generally, questions utilized close-ended binary or limited responses which allows for easier analysis of survey results.

(iv) *Was the administration of the questionnaire appropriate?*

A variety of methods including postal, electronic, telephone, and in-person interviews can be used to administer surveys. Selecting the method depends on factors such as the type of research question, design, amount and type of information needed, sample size, and available resources. The method of administration must be carefully selected while considering how it will affect the representativeness of survey respondents [8]. There may be an element of selection bias in choosing telephone surveys (e.g., respondents must be home or in an office) or electronic surveys (e.g., respondents need a computer or Internet access, which is still relevant in developing nations). In-person interviews may contribute to response bias depending on the target population and research topic (e.g., interviewer bias, sensitive subject matter).

Postal and electronic delivery methods are the most common. While electronic methods are less labor-intensive, there is a recognized trade-off that response rates remain lower [22–24]. There are strategies to improve the response rate of postal and

electronic survey questionnaires. Prenotification, an interesting survey topic, monetary incentives, personalized questionnaires, and prepaid return envelopes all enhance response rates of mailed questionnaires [13]. Each mailed reminder can yield an additional 30–50% of the initial response for both postal [25] and electronic formats [26–28]. These strategies are critical when attempting to survey physicians, a population known to have a poor response rate [29, 30]. For additional in-depth discussion on this topic, see articles by Edwards et al. [13] and Sprague et al. [31].

Lubelski et al. [5] used an online survey created using RedCAP Database, which were distributed by email. The specific timing of the survey distribution is not stated. The length of the survey and its estimated completion time are not stated. Response enhancing techniques such as incentives or follow-up reminders were not utilized.

Secondary Guides

(i) *Was pilot testing performed?*

Following pretesting revisions, pilot testing is a small-scale rehearsal of the complete survey study. The purpose is to assess the dynamics of the questionnaire and test the research process (e.g., sampling, recruitment, administration, data collection, and analysis) in field conditions [7, 19, 32]. There may be unforeseen barriers or issues to troubleshoot during the pilot study that are important to resolve prior to the proper study. Respondents are also asked to evaluate the questionnaire for its overall relevance, flow, and administrative ease [7]. Lubelski et al. [5] did not report pilot testing of their survey.

(ii) *Was clinical sensibility testing performed?*

While the focus of pretesting is primarily on the structure and understanding of individual questions, the goal of clinical sensibility testing is to evaluate the comprehensiveness, clarity, and face validity

of the questionnaire [7]. The process evaluates how well the survey assesses the pertinent clinical domains identified during development. Clinical sensibility testing should be a structured assessment (e.g., use of a standardized assessment form) by independent evaluators. Lubelski et al. [5] did not report clinical sensibility testing in their survey development. Given the complexity of the clinical vignettes, sensibility testing would have been important to ensure completeness of clinically relevant information needed for the participating surgeons to decide on their management and accurately response to the survey.

The negative impact of not performing pilot and clinical sensibility testing is demonstrated by varying degrees of missing data for survey items (ranging from 0% to 47%) [5]. According to the authors, the survey was structured so that answering survey questions were not mandatory. The rationale was to allow participants to decide on which questions to answer based on their clinical experience of managing adults, pediatric, or both populations. The consequence of this is that there is evidence of attrition with decreasing responses for each subsequent case (95 responses for case 1 pediatrics, and 81 responses for case 3 pediatrics; 115 responses for case 2 adult, and 109 responses for case 4 adult). Pilot testing may have identified this issue ahead of time. A possible solution could be restructuring the survey so that questions were mandatory to minimize missing data and gating clinical vignettes only to participants who felt comfortable answering based on the patient population they treat. Furthermore, the authors postulate that participants' perception that a question item was unnecessary or unrelated could be other reasons for incomplete data entry [5]. Such question items which may have been irrelevant to the scenario should have been identified via clinical sensibility testing and pruned.

(iii) *Was reliability and validity testing performed?*

Reliability and validity testing essentially answer the following questions:

- Does my survey answer the question consistently? (reliability)
- Does my survey answer the correct question? (validity)

Reliability determines whether questions yield consistent answers over time and repeated administration [11, 33, 34]. A reliable test can also accurately differentiate respondents [12, 33]. The response to a given question should be alike among respondents who feel similarly, and divergent among respondents who differ [12]. For interrater reliability, two independent respondents will have similar responses where expected. Furthermore, internal consistency is determined when several questions that address the same domain produce similar answers [33]. In test-retest reliability, one individual will have consistent results when answering the same question at different times [11]. In other words, the results of a reliable questionnaire are consistent and reproducible.

Validity is defined as the extent to which an instrument (e.g., questionnaire) measures what it intends to measure. Face validity is a subjective assessment of survey validity by experts and participants, often conducted during clinical sensibility testing [7, 11]. Content validity is formal expert evaluation of whether the questionnaire items accurately evaluate all aspects of the topic (i.e., all pertinent clinical domains) [7, 11]. Construct validity is determined based on a conceptual inference and relationship between the instrument and object being measured (e.g., IQ test for intelligence) [7, 11]. Lastly, criterion validity compares the results of survey items to a "gold standard" or accepted existing measures [7, 11]. Validity assessments are important for future use of a given survey in a specific group and context. However, for most surveys, only face

validity is warranted, and it is completed during clinical sensibility testing.

Overall, reliability and validity testing are important to ensure that the questionnaire can attain the true answer to a research question. This is particularly vital when questions are used to infer conclusions for a different construct based on a theoretical association or relationship. For example, a survey developed to evaluate nutritional health can ask questions on diet habits (e.g., how often do you eat vegetables?). Although there is a scientific basis for the relationship between diet and nutrition, it would be important to demonstrate that vegetable consumption is a valid measure of nutritional health through the methods described above. In contrast, face validity alone may be adequate when a survey is used for a descriptive purpose as in a cross-sectional design.

The survey by Lubelski et al. [5] does not report assessment of reliability or validity that may affect the credibility of their results. However, the reliability of the survey is likely satisfactory given the specificity of the clinical vignettes and limited-response questions. The overall descriptive objective and cross-sectional design of the survey would only require assessment of face validity, which should be performed with clinical sensibility testing.

II. What Are the Results?

(i) *Was there a sufficient response rate?*

High response rates provide confidence that the survey results are representative of the target population. This reduces potential bias between responders and nonresponders, as well as improves the validity and generalizability of the study [7]. Response rates of 60–70% are considered acceptable and demonstrate face validity in the medical community [12, 35]. Given its importance, a proper sampling frame, selected sample, and techniques to enhance the response rate are critical to achieving a high response rate.

Lubelski et al. [5] report an overall response rate of 60.4% (154/255). However, there was variable degrees of missing data to specific clinical vignettes and question items as discussed previously. If we consider each clinical vignette as a separate survey, it is unclear why there is a maximum of 115 respondents for a clinical vignette (case 2, adult) from the initial 154 surgeons that agreed to the survey. The 39 respondents that did not proceed to the clinical vignettes may represent peripheral nerve surgeons that do not manage BPI (which would underestimate the response rate), or simply lost interest and did not complete the survey. This also applies to the different response rates between pediatric and adult vignettes as surgeons may treat one or both patient populations.

Although it may have been sensible to further define the target population as peripheral nerve surgeons who treat adult or pediatric BPI, the true response rate would be indeterminable unless the organization membership registry collected such data of its members. For example, if the survey obtained 60 responses for pediatric BPI from a membership registry of 200, the true response rate cannot be determined unless we know how many members of the 200 represent the target population (i.e., those treating pediatric BPI). This information simply may not be available as the sampling frame is not precise to the target population. The result is that the chosen sampling frame likely underestimates the response rate for both pediatric and adult vignettes if they are considered separately (i.e., the inclusion of pediatric surgeons dilutes the response rate of adult surgeons and vice versa).

(ii) *Were appropriate statistical methods used?*

Surveyscan collect descriptive (raw data) or explanatory (inference between constructs) data [6, 7]. Descriptive surveys present factual data and estimate a parameter of interest. In contrast, explanatory surveys generate connections between concepts

to test a hypothesis. Surveys measuring one metric are considered unidimensional scales while those evaluating more than one metric are multidimensional scales [12]. Sample size calculation is as important for surveys as it is with other study designs, and is influenced by the research question, hypotheses, and overall design. For readers who plan to conduct a survey, we recommend consulting a biostatistician when developing your protocol.

Nonresponders and missing data can bias the results of the survey and affect its generalizability. Although strategies should be implemented during the survey design to maximize the response rate, nonresponse can be partially addressed through methods such as complete case analysis, mean imputation, and multiple imputation [36]. The appropriateness of the method chosen often depends on the mechanism of missing data (i.e., missing not at random, missing at random, or missing completely at random). The process and assumptions underpinning imputation are beyond the scope of this chapter, but readers should consider if there was any attempt made to statistically address missing data.

Lubelski et al. [5] completed a descriptive cross-sectional survey comparing practice patterns in the management of adult and pediatric BPI among international surgeons. Given the scope of the study and notwithstanding its issues with missing data, appropriate descriptive statistics are used. Additional statistical analyses were performed for categorical variables using chi-squared to test differences in decision to operate by demographic factors. Subsequently, binary logistic regression analyses to perform pairwise comparison of treatment decision if the initial test was statistically significant and reported as adjusted odds ratios (OR) with 95% confidence intervals (CI). To address missing data, the authors used a pairwise deletion method (also known as available-case analysis) to remove missing data from analysis while

utilizing all available values. The authors did not impute missing data.

(iii) *Is the reporting of the results transparent?*

The reported results of the survey should directly address the primary question. Sufficient detail is required to ensure that the findings are clear and transparent. The authors must also report methods used to handle and analyze missing data. The conclusion and implications of the study should be discussed and align with the data presented. Finally, whether the sample surveyed is representative of the population should be considered. Lubelski et al. [5] present results in absolute counts and relative frequencies, which provide sufficient raw data for reanalysis if warranted. It is explicitly stated that results were included and analyzed for individual questions even if the survey was partially completed as per the pairwise deletion approach. Access to the survey questionnaire would be valuable in an appendix to appraise individual question stems and responses.

(iv) *Are the conclusions appropriate?*

The discussion should summarize the results and implications in a concise manner that aligns with the study data. Limitations of the study, particularly with respect to methodology, should be explored. Notably, the impact of the nonresponse rate to the validity of study findings should be discussed; unlike other forms of research, responders and nonresponders often cannot be compared. Finally, surveys are self-reported and may not accurately reflect “real-world” clinical practice; results may reflect respondents’ intentions or ideal practice instead of the care practically delivered. Lubelski et al. [5] highlight several findings that demonstrate regional and specialty variations in surgical practice in the management of adult and pediatric BPI. Study limitations are appropriately discussed with respect to low response rate, risk of response bias, and missing data. Unfortunately, they do not discuss power calculation, nor the

reliability and validity of the questionnaire. The authors indicate that the results may not be generalizable but indicate a wide variety of practice patterns; this is an impetus for research and standardization in the management of BPI.

III. Will the Results Change Practice?

(i) *Are the results generalizable to my practice?*

The generalizability of the study describes how applicable the results are to the target population [6]. This is also known as the external validity. Even if a study is conducted with rigorous attention to its methodology, the results are not always generalizable. The sample being not representative of the target population is one contributing factor that is most often caused by a low response rate [7]. The survey by Lubelski et al. [5] demonstrates an overall response rate of 60.4%, but the four clinical vignettes have a response rate that range from 31.8% to 45.1%. The survey had even representation of orthopedic surgery, plastic surgery, and neurosurgery (Fig. 26.1). In contrast, regional representation demonstrates possible bias toward North American representation (55.8%) compared to Europe (24.7%), South America (10.4%), and Asia (7.1%). The low response rates combined with the uncertain selection of the sampling frame raises the question if the survey is sufficiently representative and generalizable. This impacts the validity of the differences highlighted between specialties but especially different regions of the world.

(ii) *Will the conclusions from this survey change my practice/behavior?*

The survey by Lubelski et al. [5] demonstrates statistically significant differences in surgical practice between regions and specialties. This was not consistent across different clinical vignettes. There are also concerns regarding the overall validity of the survey due to missing data and poor response rates. The findings highlight the lack of consensus among peripheral nerve surgeons in the management of BPI.

Resolution of Clinical Scenario

You are surprised by the variations in practice patterns for adult and pediatric BPI. However, you recognize some methodological weaknesses with the survey that makes you cautious of its findings. The low response rate impacts the generalizability of the results. You remain unsure regarding the degree of variation between North America and other regions of the world. You decide to present these findings at the next hand surgery rounds and identify the best-evidence underlying the practice patterns in BPI.

Conclusion

Survey research can uniquely gather information on the knowledge, beliefs, attitudes, and behaviors of a large cohort if they are properly conducted. The results of methodologically rigorous surveys can inform new trials, health policy, and effect change through knowledge translation interventions. Performing surveys can be challenging, and they must be carefully planned to obtain meaningful and unbiased results. We recommend consulting a biostatistician when developing the study design and methodology. This chapter demonstrates the importance of survey design, appropriate testing of questionnaires, methods of enhancing response rates, appropriate analysis of study results, and their interpretation.

References

1. Alderman AK, Hawley ST, Waljee J, Morrow M, Katz SJ. Correlates of referral practices of general surgeons to plastic surgeons for mastectomy reconstruction. *Cancer* [Internet]. 2007;109(9):1715–20. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17387715>
2. Mathes DW, Schlenker R, Ploplys E, Vedder N. A survey of north american hand surgeons on their current attitudes toward hand transplantation. *J Hand Surg Am* [Internet]. 2009;34(5):808–14. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19410983>
3. Coroneos CJ, Roth-Albin K, Rai AS, Rai AS, Voineskos SH, Brouwers MC, Avram R, Heller B. Barriers, beliefs and practice patterns for breast cancer reconstruction: a provincial survey. *Breast*

- [Internet]. 2017;32:60–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28038321>
4. Wingrat J, Berger C, Bauer L. Passive range of motion for conservative management of brachial plexus birth injury: a survey of expert therapists. *Occup Ther Health Care* [Internet]. 2023;29:1–10. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/38019167>.
 5. Lubelski D, Feghali J, Hersh A, Kopparapu S, Al-Mistarehi A, Belzberg AJ. Differences in the surgical treatment of adult and pediatric brachial plexus injuries among peripheral nerve surgeons. *Clin Neurol Neurosurg* [Internet]. 2023;228:107686. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/36963285>
 6. Thoma A, Cornacchi SD, Farrokhyar F, Bhandari M, Goldsmith CH. Evidence-based surgery working group. How to assess a survey in surgery. *Can J Surg* [Internet]. 2011;54(6):394–402. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21939608>.
 7. Burns KEA, Duffett M, Kho ME, Meade MO, Adhikari NK, Sinuff T, Cook DJ, ACCADEMY Group. A guide for the design and conduct of self-administered surveys of clinicians. *CMAJ* [Internet]. 2008;179(3):245–52. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18663204>.
 8. Burns KEA, Kho ME. How to assess a survey report: a guide for readers and peer reviewers. *CMAJ* [Internet]. 2015;187(6):E198–205. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25691790>
 9. Rubenfeld GD. Surveys: an introduction. *Respir Care* [Internet]. 2004;49(10):1181–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15447800>
 10. Aday LA, Cornelius LJ. Designing and conducting health surveys: a comprehensive guide. 3rd ed. San Francisco: Jossey-Bass; 2006.
 11. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Designing clinical research. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2013.
 12. Passmore C, Dobbie AE, Parchman M, Tysinger J. Guidelines for constructing a survey. *Fam Med* [Internet]. 2002;34(4):281–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12017142>
 13. Edwards PJ, Roberts I, Clarke MJ, Diguiseppi C, Wentz R, Kwan I, Cooper R, Felix LM, Pratap S. Methods to increase response to postal and electronic questionnaires. *Cochrane Database Syst Rev* [Internet]. 2009;(3):MR000008. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19588449>.
 14. Stone DH. Design a questionnaire. *BMJ* [Internet]. 1993;307(6914):1264–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/8281062>
 15. Jones TL, Baxter MAJ, Khanduja V. A quick guide to survey research. *Ann R Coll Surg Engl* [Internet]. 2013;95(1):5–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23317709>
 16. Babbie E. Survey research methods. 2nd ed. Belmont: Wadsworth; 1973.
 17. Fox J. Designing research: basics of survey construction. *Minim Invasive Surg Nurs* [Internet]. 1994;8(2):77–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/7812386>
 18. Collins D. Pretesting survey instruments: an overview of cognitive methods. *Qual Life Res* [Internet]. 2003;12(3):229–38. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12769135>
 19. Ruel E, Wagner WEJR, Gillespie BJ. The practice of survey research: theory and applications. Thousand Oaks: SAGE Publishing; 2015.
 20. Williamson PR, Altman DG, Bagley H, Barnes KL, Blazeby JM, Brookes ST, Clarke M, Gargon E, Gorst S, Harman N, Kirkham JJ, McNair A, Prinsen CAC, Schmitt J, Terwee CB, Young B. The COMET handbook: version 1.0. *Trials* [Internet]. 2017;18(Suppl 3):280. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28681707>.
 21. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* [Internet]. 2010;19(4):539–49. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20169472>
 22. Kim HL, Hollowell CM, Patel RV, Bales GT, Clayman RV, Gerber GS. Use of new technology in endourology and laparoscopy by American urologists: internet and postal survey. *Urology* [Internet]. 2000;56(5):760–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11068295>
 23. Raziano DB, Jayadevappa R, Valenzuela D, Weiner M, Lavizzo-Mourey R. E-mail versus conventional postal mail survey of geriatric chiefs. *Gerontologist* [Internet]. 2001;41(6):799–804. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11723348>
 24. Braithwaite D, Emery J, De Lusignan S, Sutton S. Using the Internet to conduct surveys of health professionals: a valid alternative? *Fam Pract* [Internet]. 2003;20(5):545–51. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/14507796>
 25. Sierles FS. How to do research with self-administered surveys. *Acad Psychiatry* [Internet]. 2003;27(2):104–13. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12824111>
 26. Fischbacher C, Chappel D, Edwards R, Summerton N. Health surveys via the Internet: quick and dirty or rapid and robust? *J R Soc Med* [Internet]. 2000;93(7):356–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/10928022>
 27. McLean SA, Feldman JA. The impact of changes in HCFA documentation requirements on academic emergency medicine: results of a physician survey. *Acad Emerg Med* [Internet]. 2001;8(9):880–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11535480>
 28. Schleyer TK, Forrest JL. Methods for the design and administration of web-based surveys. *J Am Med Inform Assoc* [Internet]. 2000;7(4):416–25. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/10887169>

29. Asch DA, Jedrzewski MK, Christakis NA. Response rates to mail surveys published in medical journals. *J Clin Epidemiol* [Internet]. 1997;50(10):1129–36. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/9368521>
30. Cummings SM, Savitz LA, Konrad TR. Reported response rates to mailed physician questionnaires. *Health Serv Res* [Internet]. 2001;35(6):1347–55. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11221823>.
31. Sprague S, Quigley L, Bhandari M. Survey design in orthopaedic surgery: getting surgeons to respond. *J Bone Joint Surg Am* [Internet]. 2009;91(Suppl 3):27–34. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19411497>.
32. Bowden A, Fox-Rushby JA, Nyandieka L, Wanjau J. Methods for pre-testing and piloting survey questions: illustrations from the KENQOL survey of health-related quality of life. *Health Policy Plan* [Internet]. 2002;17(3):322–30. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12135999>.
33. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 4th ed. Oxford: Oxford University Press; 2008.
34. Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis* [Internet]. 1985;38(1):27–36. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3972947>.
35. Elder JP, Artz LM, Beaudin P, Carleton RA, Lasater TM, Peterson G, Rodrigues A, Guadagnoli E, Velicer WF. Multivariate evaluation of health attitudes and behaviors: development and validation of a method for health promotion research. *Prev Med (Baltim)* [Internet]. 1985;14(1):34–54. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/4034513>.
36. Rubin DB. *Multiple imputation of nonresponse in surveys*. Toronto: Wiley; 1987.



Simple Statistical Tests and P Values

27

Charles H. Goldsmith, Eric Duku, Achilles Thoma,
Jessica Murphy, and Emily Dunn

Abbreviations

ASA	American Society of Anesthesiologists
PICOT format	P = population, I = intervention, C = comparative intervention, O = outcome, T = time
RCT	Randomized controlled trial
CI	Confidence interval
SD	Standard deviation
DF	Degrees of freedom
MID	Minimum important difference
ITT	Intention-to-treat
RR	Relative risk

Clinical Scenario

A 75-year-old retired surgeon is seen in your office after a recent colonoscopy identified a malignant polyp in his colon. As a colorectal surgeon, you recommend a segmental colon resection. You are concerned however by his overall poor physical condition. He is overweight, has hypertension, and more importantly has dyspnea. His medical history notes that he was a former smoker for 30 years. You guess that he has a score III/IV on the American Society of Anesthesiologists (ASA) scale. You inform him that you will be asking for an anesthesia consult

C. H. Goldsmith (✉)
Department of Health Research Methods, Evidence
and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: goldsmi@mcmaster.ca

E. Duku
Department of Health Research Methods, Evidence,
and Impact, McMaster University,
Hamilton, ON, Canada

Offord Centre for Child Studies, Department of
Psychiatry and Behavioral Neurosciences,
McMaster University, Hamilton, ON, Canada
e-mail: duku@mcmaster.ca

A. Thoma
Division of Plastic Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: athoma@mcmaster.ca

J. Murphy
Division of Physical Medicine & Rehabilitation,
Department of Medicine, McMaster University,
Hamilton, ON, Canada
e-mail: murphj11@mcmaster.ca

E. Dunn
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: dunne4@mcmaster.ca

before surgery. He asks you if you would consider a prehabilitation program for him before surgery. You inform him that you are not certain if this helps but you are willing to consider it. You promise to check if there is any evidence to support this approach. You are very skeptical of published studies that proclaim the superiority of some novel intervention and show a P value to justify their claim.

Literature Search

As stated in previous chapters, the first step to finding the best available evidence is to formulate a research question based on the PICOT format:

- Population: High-risk patients for abdominal surgery
- Intervention: Prehabilitation Intervention
- Comparative intervention: Standard care
- Outcome: Complications
- Time: Any time after surgery

Using the PICOT format terms, your clinical research question is: In high-risk patients who undergo abdominal surgery, does a prehabilitation intervention, compared to standard care, reduce the chance of postoperative complications any time after surgery?

To answer your research question, you conduct a literature search using PubMed. From the PICOT format terms mentioned above, you choose the following for your search: High Risk AND Abdominal Surgery AND Prehabilitation AND Complications; you narrow your search to the last 10 years (2013–2024). Your search yields 13 articles (See Appendix A). You screen the articles first by title, then by abstracts, and find several promising articles including some meta-analyses. However, following this screening, the article by Barberan-Garcia et al. [1] appears to be the best suited for the research question and the purpose of this chapter, which is to review introductory statistical methods in-depth. This article focuses on a personalized prehabilitation program for high-risk patients. It was published in 2018 and it utilizes a randomized,

blinded controlled study design. The other articles found in your search focus on general and digestive surgery populations, sarcopenia, or are systematic reviews, feasibility studies, case reports; because of this, you are confident in using the Barberan-Garcia et al. [1] article.

Chapter Goal

The authors of this chapter would recommend that readers consider reading the Chap. 29: *Power and Sample Size* as some of the overlapping ideas are discussed there.

Reporting of studies in the surgical literature usually contains the use of statistical principles such as estimation and hypothesis testing. These ideas are often displayed using equations and mathematically oriented ways of doing calculations with a calculator; however, statistical software has been widely developed to do these calculations. The goal of this chapter is to describe the statistical principles in English with accessible language, leaving the mathematical symbolism to a minimum. We have decided to keep our estimation and hypothesis testing to a few simple methods, which are often seen in the surgical literature. We leave the detailed statistical and mathematical descriptions to the statistics literature. Some of the references listed throughout the chapter have more detailed discussions of these statistical ideas to make them more explicit for the mathematically inclined reader.

The clinical article that the authors have selected to illustrate the statistical principles is a randomized controlled trial (RCT) with two arms (or groups): control and intervention. The primary outcome measure is complication rates in the two groups. Two secondary outcome measures are stated as the number of complications per patient and severity of these complications. The paper by Barberan-Garcia et al. [1] reports other findings using other statistical tests and P Values; however, these are not described in this chapter. Our discussion will focus on those used in comparing these two groups with descriptive statistics, test statistics with their names, and how they can be computed with reliable statistical

software. We will then reproduce these statistical test results and compare our results with those reported in the Barberan-Garcia et al. article [1]. Along the way, we will comment on the assumptions that go with these tests. We do not produce tests of these assumptions, just whether the paper’s authors [1] have provided enough detail to convince us that the tests used and conclusions drawn from them support the findings we need to resolve the clinical scenario. We include confidence intervals (CIs) and P values to help interpret the findings in the appraised paper [1]. [For more information, readers may consult Chap. 28: *Confidence Intervals* or work by Cadeddu et al. [2]]. Numbers, tables, and text will be used to describe the statistical output from the cited software as well as how the software does the computations. Graphs, although they often can be helpful to our understanding of these ideas, will not be used.

Critical Evaluation of the Paper for Reporting RCTs

This chapter uses the validity checklist for scoring meta-analysis articles by Cochrane [3] as it seemed to be most relevant to this scenario (see

Appendix C). The interventions in our scenario are not surgical; they compare prehabilitation to no prehabilitation (standard care), which seem to be used with surgery, however, are mostly exercise related and not surgical.

The Risk of Bias tool was scored by CHG. Of the items in the list, one [Randomization] was scored Yes, indicating it was done well. However, the rest were either scored No or Unsure, which usually gets interpreted as a No. So, in summary, Barberan-Garcia et al. [1] article was Considered Flawed, as it was not reported up to the standards usually seen in the RCT literature.

From the information in Table 27.1, there are some different statistical tests and CIs used that we will discuss. Then we will use software to recompute the findings in the Barberan-Garcia et al. [1] article. These recomputed findings can be compared to what the Barberan-Garcia et al. [1] article stated. Last, we will look at the impact of the conclusion for our clinical scenario. Along the way, we will refer to available resources, which can be found in the reference list that may provide readers with more details of the statistical issues and may be helpful to readers who like additional information or want another viewpoint than we present.

Table 27.1 Summary of findings in the paper [1] that use text to display things relevant to the outcome

Study element	Information	Pg.
Study objectives		
Main outcome	Proportion of patients suffering postoperative complications (any deviation from the normal postoperative course)	50
Secondary outcomes	Mean number of complications per patient and severity of postoperative complications	52
Methods		
Sample size justification	Complication rate of similar patients: 30% α -Risk of 0.05 β -Risk of 0.2 2-sided test Anticipated 20% dropout rate Minimal clinically important difference: Decrease in percentage of patients with complication (intervention vs. control) of 20% or more <i>Total needed: 70 participants per group</i>	52
Excluded patients	19: Due to the changes in surgical plan	50
Dropout rate	19/144 = 13.2%; (less than the 20% used in the sample size justification)	52

(continued)

Table 27.1 (continued)

Study element	Information	Pg.
Randomization	71 in control arm; 73 in intervention arm	50
Modified intention-to-treat analysis	63 from control; 62 from intervention group were included	50
Results		
Statistical tests	Categorical variables: Chi-square or fisher exact tests. Numerical variables: Student's t or Wilcoxon tests	52
Presentation of data	Mean and standard deviations, or count and percentages, as indicated	52
Baseline characteristics	Characteristics were balanced, including: Complexity of the surgical approach, in the modified intention-to-treat patients Intraoperative variables	52
By group	The intervention group had a lower rate of complication (31%) as compared to the control group (62%) ($p = 0.001$)	53
Relative risk (RR)	Estimated RR of 0.5 (CI 0.3–0.8) for complications suggested that prehabilitation has a protective role for postoperative complications	53
Mean (SD) by group	The intervention group had a lower complication mean (SD) 0.5(1.0) as compared to the control group 1.4(1.6); ($p = 0.001$)	55
Severity	No effects on severity of complications, as measured by the Clavien-Dindo classification, (data not shown)	53
Conclusion		
Prehabilitation	Prehabilitation enhanced clinical outcomes following surgery	50
Complications	The incidence of complications in patients was 46%	53

These findings in Table 27.1 are taken from the Barberan-Garcia et al. [1] paper; the results have been organized by study element and page number where this information can be found

CI Confidence Interval, SD Standard Deviation

Statistical Tests and P Values

In Table 27.2, we list the Test Statistics, the Null Hypothesis, and their Reference Distributions for computing the associated P values that will be encountered in the resolution of the clinical scenario. These tests can also be analyzed using confidence intervals to test the null hypotheses listed. Both Chi-square and Student's t Reference Distribution need to know the degrees of freedom (DF) from a test statistic for a P value computation. The Fisher Exact test provides a way to get the P value by direct computation from a Table of counts like Table 27.3 without creating any intermediate numerical test statistic. While the statistical literature contains many thousands of different statistical tests, we will keep to those that are needed to be able to understand and interpret the test statistics and their P values to compare with the output from the Barberan-Garcia et al. paper [1].

In Table 27.1, it is clear that a Chi-square test was done to compare the proportions (or percent-

ages) between the trial arms as well as the relative risk comparing the two groups, although which test was used to report the findings was NOT reported in each of the result tables. This choice came from the statistical software (cited later when used) that was used to recompute the findings that are reported. Each test statistic has a corresponding confidence interval (CI), although just the 95% CI was reported in the paper. These CIs were recomputed with the statistical software, although the confidence coefficient (the 95) usually can be changed in many statistical software packages. The criterion used to judge statistical significance (the level of statistical significance) was set at $\alpha = 0.05$ in Table 27.1. The confidence coefficient is computed from the α value by determining its complement and expressing it as a percentage, or $(1-\alpha)100\%$, in this case, 95%. From Table 27.1, we also saw that the authors reported two-sided tests; this also means that the confidence intervals are two tailed, showing a minimum and a maximum as two numbers, usually reported as (minimum to maximum) to avoid using a dash (–) to

Table 27.2 Test Statistics, null hypotheses and reference distributions

Test statistic	Null hypothesis	Reference distribution
Chi-square	Statistical independence	Chi-square
Two independent sample Student's t	0(zero)	Student's t
Two independent sample proportions	0(zero)	Normal (0,1)
Relative risk	1(one)	Normal (0,1) on the log _e scale
Fisher exact	Statistical independence	Hypergeometric

Table 27.3 Primary outcome counts with modified ITT analysis

Complications\Arm	I	C	Total
Yes	19	39	58
No	43	24	67
Total	62	63	125

Table 27.4 Numerical Chi-square test results

Tests/values	Chi-square	DF	P Value
Pearson	12.277	1	0.000
Likelihood ratio	12.495	1	0.000

separate them if the data could be reported as negative numbers, such as with differences.

The Chi-square test can be computed from a 2×2 contingency table illustrated in Table 27.3. The rows in the table designate complication counts (Yes) and non-complication counts (No) with the bottom margin of the table showing Total Counts for each column. The first two Total Counts in the last row come from the header in Table 27.4 of the Barberan-Garcia et al. [1] article. The columns in the table show I (Intervention) counts, C (Control) counts, as well as Total counts for each row of the table.

The Yes row was derived from Table 27.4 of the appraised article [1], the Total by adding $19 + 39 = 58$. The Total in the bottom right corner is obtained by adding the two sample sizes $62 + 63 = 125$, the number of patients that Barberan-Garcia et al. [1] claim is the Intention-To-Treat (ITT) total. However, the paper's authors [1] have misused this term; it should be "Modified ITT" as it leaves out the 19 patients that were randomized, Table 27.1, but did not get surgery. We are NOT told details of these 19 patients as they were also left out of the Table 27.1 in the original article [1], and no complications were recorded or presented for them, even though the Intervention patients could have had compli-

cations resulting from the Prehabilitation Intervention, and the Control patients may have had complications in any event.

Using the data in Table 27.3, the overall rate of complications is $58/125 = 0.464$ or 46.4%, which the authors report as 46% but not as a proportion, Table 27.1. [While it is simple to convert proportions to percentages and vice versa, they can matter when you enter them into statistical software when one or the other is specifically required.] For the Control group, the rate of complications is $39/63 = 0.619$ or 61.9%; for the Intervention group the rate of complications is $19/62 = 0.306$ or 30.6%. The paper authors reported these as integer percentages: 62% and 31% (Table 27.1). [Reporting to more decimal places is not reasonable for these sample sizes; however, the extra decimal places should be used to do further calculations such as with a relative risk. We use the statistical software to compute these here.]

Could All of This Be Due to the Play of Chance?

To consider the role of chance in our results, the sample size calculation, Table 27.1, used by Barberan-Garcia et al. [1] is based on percent-

ages, not proportions (Table 27.1). A background complication rate of 30% using colorectal patients, which could be considered as the Control Group rate, was used. Also, the reduction in the rate was stated to be at least 20% and so this implied either a difference in percentages of that size to be considered as a clinically important Minimum Important Difference (MID) [4], with similar interpretations for ratios such as the relative risk. From proportions (percentages), the variance and standard deviation can be computed, so they are not needed to be stated separately. Also, α and β are defined, so all 4 Greek letters: α , β , δ , σ are specified along with the two-sided or two-tailed tests [Ts] (Table 27.1), and the study design (D). While the software used to calculate the sample size needed for the trial is specified, the outcome measure [O] and test statistic [Tt] were not specified among the four tests that Barberan-Garcia et al. [1] planned to use (Table 27.1). Therefore, six of the eight criteria for a sample size calculation (Recall $n = f(\alpha, \beta, \delta, \sigma, D, O, Ts, Tt)$ from Chap. 29: *Power and Sample Size*) were specified by the paper's authors [1]. The paper's authors also inflated the computer output to accommodate 20% dropouts, to get the 70 needed per group. [Of the 140 required, the authors randomized 144.] For our discussion here, we assume that the outcome (primary) was the rate of complications and the test statistic was the Chi-square test.

Computing the Test Statistic

The data in Table 27.3 were entered into Minitab 18 [5] using the "Chi-square Test for Association." The output is shown in Table 27.4. [We do not show here that the output included the data in Table 27.3, which is useful to check that the data entry step was correct.]

For this discussion, we will use the Pearson method for computing the value of the Chi-Square test statistic. [If there is a choice in statistical software output, surgical researchers should state which one was chosen and why.] In this case, the value of the test statistic is 12.277, it was computed with 1 DF and the P value was

0.000. [It is important that researchers be careful with statistical software output. The number of decimal places printed can be modified in most software packages. Minitab 18 [5] produces three decimal places by default. This does not mean that you should blindly report this P Value in your paper; it should be reported as "< 0.001" (Table 27.1). If you are interested in the other method for computing the Chi-square test statistic, Minitab 18 [5] has a help file to explain the differences between the two methods.

What Does the Number Mean to a Reader?

For the data in Table 27.3, when we compute a Chi-square test statistic, this is the value of the number used to judge whether the null hypothesis being tested is true. Association in the null hypothesis is statistical independence between the arms of the trial as columns in the table and the rows that have the numerators/denominators for the percentages (or proportions) of complications, such as in Table 27.3. For a test statistic with 1 DF, we expect numerical values in the region of the mean = 1, so the 12.277 seems bigger than the mean; it is in the right-hand tail of what we would expect.

We can also use the data in Table 27.3 to do a statistical test that compares two proportions (or percentages) where we compute a difference between the two proportions as the test statistic and its associated CI. Here the null hypothesis is that the population difference between the two arms is 0 (zero) and the test statistic has a sampling distribution of the test statistic that is normal with a mean of 0 (zero) and variance (and standard deviation) of 1(one). However, since differences can be negative or positive, the sampling distribution has two tails that we could decide to test a one tail test to get a P value as 1P or as P1, depending on which tail is to be set by the alternative hypothesis. We will use both tails because we would like to detect either harm or benefit for the intervention of prehabilitation; i.e., compute P2. The sample size justification also appears to be claiming a MID of a least a 20% difference in complications rates to be credible as

well as detectable in the trial justification (Table 27.1). Because Barberan-Garcia et al. [1] did not mention which test statistic or the outcome measure explicitly, one might surmise that the paper's authors meant the testing of equal proportions (percentages) of complications from the two arms of the trial.

Where Do *P* Values Come from?

For each test statistic, if we were to repeatedly sample, say, hundreds of times, the data from the study as we have here [Nobody ever does this in reality; however, that is what the statistical theory postulates], a two-arm trial on the prehabilitation intervention, there would be hundreds of different numbers for that test statistic in these samples. For a large number of samples, statistical theory tells us the distribution of these Chi-square test statistics has a Chi-square distribution with 1 DF. This distribution is sometimes called a reference distribution as *P* values are computed from this reference distribution considering the numerical value that corresponds to the significance level specified for the study; here $\alpha = 0.05$. With the Chi-square test statistic, both positive and negative deviations are computed from what would be expected by the null hypothesis. The “square” part of the Chi-square test statistic indeed squares computationally these numbers so both the large negative deviations and large positive deviations from the null hypothesis get relegated to the right-hand tail of the reference distribution. There are two tails to the set of test statistics, but they get forced into a single right-hand tail of this reference distribution. Even so, from the original data there are two tails involved so this becomes a two tail test, as required by Barberan-Garcia et al. [1] and for our clinical scenario. One way to see this is to have stated this as *P*₂ to indicate a two tail test, *P*₁ as a one tail test in the right-hand part of the departure from the null hypothesis and *P*₁ as a one tail test in the left-hand part of the departure from the null hypothesis. So, in this case, we could just as easily report that $P_2 < 0.001$, rather than the notation that the paper's authors used.

The Chi-square reference distribution with 1 DF has a mean of 1, variance of 2 and standard deviation as the square root of 2 or 1.414 to 3 dp (decimal place). So the computed test statistic of 12.277 seems to be many standard deviations $[(12.277-1)/1.414 = 7.97]$, around 8] above the mean in the reference Chi-square distribution. To make this more specific, given the size of level of significance of 0.05 and the area under the reference distribution of 1, the chance or probability that the reference distribution exceeds 12.277 will be the area in the right-hand tail. Minitab 18 [5] has computed this as 0.000 in Table 27.4 and it should be reported as $P_2 < 0.001$. Because this *P*₂ value is less than the significance level, these complication rate data are declared to be statistically significant.

Many times there are standard assumptions that need to be met if the sampling distribution for the test statistic is well described by the reference distribution. These should be reported as being checked in any paper using these distributions. For example, the assumptions for the Chi-square test statistic to have a Chi-square reference distribution for *P* value calculation are: Two independent arms, each patient in each arm should be independent of each other, and the allocation of patients to arms should be random and have complete reporting. The first three seem to be satisfied; however, not the last one, as the authors use a Modified ITT, leaving out those 19 patients who were randomized but did not get surgery (Table 27.1). A sensitivity analysis could be used in this case to compare the results of those randomized to those used in the modified ITT analysis; which we report later. Multiple imputation [6] is not sensible here because the baseline variables for the 19 patients left out from the results were not available to be used for the imputations.

There is currently much controversy about using *P* values to judge the importance of research findings. Some of the debate is cited in the references [7–13]. Some of the remedies to help include using CIs to express the study findings. Please see Chap. 28: *Confidence Intervals*, for more information.

A second way of reporting the findings in Table 27.3 is to use a null hypothesis related to

relative risk (RR) of complications in the Intervention group compared to the Control group. An RR compares the rate of complications in the Intervention group to the rate in the control group using the ratio with Intervention in the numerator and Control in the denominator. The null hypothesis here is that the population RR is 1(one). [Be careful with the data entry into the software used as the reverse for the numerator and denominator can give the inverse of RR as 1/RR instead.] To do the computation here we are using statistical software called CIA for Confidence Interval Analysis [14] to do the computations with confidence intervals rather than P values. Using the 2×2 entries in Table 27.3, we use: “Relative Risk for parallel groups,” with results are shown in Table 27.5.

To report these values like in the Barberan-Garcia et al. [1] paper (Table 27.1), the estimate and the boundaries of the 95% CI are rounded to 1 dp as 0.5; 0.3 to 0.8. Using the output we could choose normal approximation of the logarithms of the RR null hypothesis test and this can be reported $P2 < 0.001$ for it or $P2 = 0.001$ for the Fisher Exact test. This compares favorably with Table 27.1, except for which test was used. Our computations would agree with the Barberan-Garcia et al. [1] using the two proportions test statistic, but NOT with the Normal approximation; since the Fisher Exact test was not needed. The event rates were not rare or highly skewed

where the Fisher Exact test gets used. The reference distribution for the Fisher Exact test is the hypergeometric distribution. The assumptions for the relative risk tests are the same as for the Chi-square test; however, the normal approximations for the test statistic depends on the proportions being between 0.2 and 0.8 as they were here. Otherwise, the values could be highly dependent on the distance away from 0.5 in the tails of the proportion distribution. Sensitivity analysis shown in Appendix B for the Chi-square test should be similar for the comparison of the ratio of proportions called RR as well as the Fisher Exact test. The one corner shown to be bigger than $\alpha = 0.05$ and so not statistically significant should be the same as the RR in Appendix B, although we do not show these.

Now, what about the secondary outcomes computed from the complications; their mean and standard deviations with a P value is shown in Table 27.6. These are called rates and are computed from the total count of complications divided by the sample size values per patient rather than whether or not a patient had at least one complication, the Yes row in Table 27.3. They could then be tested with a two independent sample Student’s t test. Computations are done in Minitab 18 [5]. However, the SDs are both bigger than the means and suggests that the distribution of the number of complications is skewed right. It is also possible that the SD for the control group

Table 27.5 Proportions for relative risk calculations with Minitab [5]

Statistic	Estimate	95% CI	Minitab P+	Fisher Exact P+
RR	0.495	0.325 to 0.755	0.000	0.001

Note: + Computed with proportions test in Minitab 18 [5], reported as $P2 < 0.001$ for the Normal Test of the difference in proportions as well as the Fisher Exact test as $P2 = 0.001$. However, the latter is not usually used since all of the counts in Table 27.3 exceed 5

Table 27.6 Secondary Outcome: Results per patient by group using Minitab [5]

Group	n	mean	SD	SE mean++
Control	63	1.40	1.60	0.20
Intervention	62	0.50	1.00	0.13
Total+	125			

Note: + Added to the table by CHG with calculator computation with these entries. ++ is added by Minitab 18 [5], not in Barberan-Garcia et al. [1]. The rest of the results may have been rounded to 1 dp by Barberan-Garcia et al. [1]; the raw data were not in the appraised article [1] to check their veracity. Having a distribution of the counts of complications for all patients by group would have permitted their recomputation

is larger than the intervention group so might violate the variance homogeneity assumption of the usual Student's *t* test. We may have to use the Welsh variation of the Students' *t* test, which does not assume variance homogeneity. [CHG's rule without testing variance homogeneity is doubling of one SD over the other, which is not true here.] The assumptions of randomness is still a problem, but independence between and within the groups is still satisfied.

The equal variance assumption for the Student's *t* test computed the mean difference as 0.900, pooled SD of 1.337 with a 95% CI: 0.427 to 1.373. Since the null hypothesis here would be 0 (zero), it is clear that the 95% CI does not contain the null value. The pooled Student's *t* test is 3.76 with 123 DF and a *P* value of 0.000. This could be reported as $P_2 < 0.001$ and is lower than the significance level of $\alpha = 0.05$, so the results are statistically significant. We were not given what Barberan-Garcia et al. [1] consider to be the MID for this measure, so the clinical meaning may not be possible without other data.

The Welsh *t* test does not assume that the variances or SDs of the two groups are equal, and so Minitab 18 [5] does not report the pooled standard deviation in its output. The mean difference is still 0.900, with a 95% CI: 0.428 to 1.372. Since the null hypothesis here would also be 0 (zero), it is clear that the 95% CI does not contain the null value. The Welsh *t* test is 3.78 with 104 DF and a *P* value of 0.000. This could be reported as $P_2 < 0.001$ and is lower than the significance level of $\alpha = 0.05$, so the results are statistically significant.

The impact of the variance homogeneity assumption is small. The 95% CIs are different in the third dp, and test statistics are different by 0.02, a small amount and the DF are quite different suggesting that the reference Student's *t* distribution for computing the *P* value is different; however, to 3 dp, the *P* values are both 0.000 and regardless of which one was chosen it still would be reported a $P_2 < 0.001$. Hence the variance homogeneity assumption would not make any difference to the interpretation of the tests.

Like the other tests, the two independent sample Student's *t* test would have been nice to

be able to do a sensitivity analysis to counter the Modified ITT that was reported in the appraised article [1]. The one thing that we could do is to assume that all data from the 19 patients left out of the analysis had some specific number of complications, which were not reported.

Last, the third way to consider the complications was to analyze the severity data on the complications. However, we were not given any data on the severity except how they were coded with no reference citation of how. Barberan-Garcia et al. [1] claimed that there was no difference in the severity (Table 27.1). However, leaving out these 19 patients, this claim does seem credible given the large differences shown in enhanced aerobic capacity with ΔET , $P < 0.001$ and the reduced length of stay in the ICU shown in Table 27.4 of the appraised article [1], even though $P = 0.078$. Data should have been provided to support this finding.

Confidence Intervals

Just like hypothesis tests and *p* values, confidence intervals can be one or two tailed. Our discussion here uses two-tailed confidence intervals, although Minitab 18 [5] and CIA [14] allow one-tailed confidence intervals if they make sense. Readers are advised to read Streiner's article [15] for a discussion of the reasoning for the choices between one- and two-tailed tests.

Software

Below are three commonly used, and easily accessible statistical software options that readers may consider utilizing.

R: This software is free and is supported by statisticians all over the world under the creative commons license. As this chapter is being written it has survived for almost 30 years, and as of November 2024, currently contains 21,738 available packages [16]. All of these are available for anyone to try for their personal and study usage. Package authors tend to be very responsive to sensible questions. Readers wanting more information are invited to consider the references [6, 21–23].

CIA: This software is specifically for computation of confidence intervals. [14] It comes as part of the book when purchased.

Minitab [5]: The current version of this software is 21.1. Academic and student versions available for those with appropriate academic affiliations.

Those who choose to use software such as Microsoft Excel, web, cell phone, and cloud apps for their statistical calculations should be wary of their poor computational validity as suggested by many different peer-reviewed publications [17–20]. Your collaborative biostatistician should help you find the reviews of software in these as different new devices and locations that you can cite when you use such questionable sources to support your surgical research.

Sensitivity Analysis

A credible RCT needs to report all the data for all patients randomized to conduct an ITT analysis, which is considered the Gold Standard way to report study findings. When that does not happen as in the Barberan-Garcia et al. [1] paper, variations from ITT such as a Modified ITT reported in the appraised paper [1] need to be supplemented with a sensitivity analysis for any assumption that has not been met to make the statistical analysis valid. We conducted such a sensitivity analysis by comparing the Modified ITT reported in the appraised paper [1] and making some assumptions about the 19 patients randomized but removed from the tables in the Barberan-Garcia et al. [1] paper. For these 19 patients, we are not given any baseline data on them in Table 27.1 of the original paper [1], so we do not know any of their baseline characteristics. This prevented us from doing multiple imputations of these data related to these 19 patients to see if it had any impact on the study conclusions. Such imputations can be conducted using R software [6, 21–23] and should be used to study the missing data impact on the conclusions as outlined in the New England Journal of Medicine [22] that summarizes a 2010 National Science Foundation report. A NEJM editorial published with the article suggests NEJM does not want to see manu-

scripts if this is not done; the surgical literature should take heed.

We have created a sensitivity analysis shown in Appendix B that measures the veracity of the group difference analysis done with the primary outcome of Rates of Complications as they are reported in Table 27.3. To do so, we create four different uses of whether the 19 patients would have an impact on the analysis and reporting by including these patients as having had complication (Yes or No) between the two arms of the trial data, making tables that contain all 144 patients randomized versus Table 27.3, which has 125 and excludes these 19 patients. Notice that all the created tables have a total of 144 patients, with 73 in the intervention group and 71 in the control group, exactly as randomized. The P values created by the Chi-square tests as done for Table 27.3, show that three of these analyses have P values that are below the level of significance, so these sensitivity analyses also show statistically significant findings, and effect sizes denoted by Δ that are larger than the MID of 20%. However, one of the sensitivity analyses, *c*, showed P values that are larger than the significance level, and the Δ of 14% is smaller than the stated 20% so this case is NOT statistically significant, nor a clinically important finding.

In summary, three of the four sensitivity analyses support the Modified ITT analysis and one of the four does not. The dashes in the *e* part of the tables were not computed, yet in the right-hand corner of the 3×3 tables, there may be other cases apart from *c* where the sensitivity data do not support the Modified ITT analysis. This suggests that the fact of the discrepancy among these sensitivity analyses created by the missing 19 patients casts serious doubt about the clinical importance as well as the statistical significance of the prehabilitation reported in the Barberan-Garcia et al. paper [1]. Although we do not show it here, we anticipate that the relative risk analysis and the proportions analyses would also NOT pass the sensitivity analyses of these assumptions using tables similar to those in Appendix B; however using these other test statistics.

If we had access to the Table 27.1 data at baseline for the 19 deleted patients, a strategy called multiple imputation analysis could be conducted with the package called *mice* (multiple imputation using chained equations) in R as outlined by van Buuren [6]. What imputation does is produce models that are fitted to the available data using variables in the study to make multiple predictions for all the missing data. Next, for each imputed data set, do a complete planned analysis as described in a statistical analysis plan for the study, to measure the impact of missingness on the conclusions and this impact should be reported as part of the Barberan-Garcia et al. paper [1], as now some journals require [24].

Multiplicity

When there is more than one source of data that reflect on a single outcome variable such as the Barberan-Garcia et al. paper [1] with complications measured in three ways: counts, mean per patient, and severity, it means a reader trying to decide about a clinical scenario needs to consider which of the three to use in the resolution of the scenario. We already know that the paper's authors claimed that the Counts was considered to be the primary outcome measure and presumably that is what was used in the sample size justification, although not stated by the appraised paper authors [1]. What other choices could be made using this multiplicity of outcome measures. Multiplicity leads to consideration of how one might use the risk of a Type I error called α , which we now know drives the P values to determine whether the test statistics used are statistically significant. One common way is to consider the three ways judged separately with the Bonferroni adjustment of the level used by having each of the outcomes compared to $\alpha/3 = 0.05/3 = 0.0266$. With our scenario, this will not work since the adjusted level has an assumption of statistical independence, which is not true since all three ways are derived from the same outcome and hence are highly correlated. Because of this, we chose NOT to make any level adjustment for the interpretation of the test statistics. If all three of these ways are to be judged simultaneously, then the proper way to do so is

with a Multivariate test statistic, called Hotelling's T Squared test; however, we could NOT consider doing this due to lack of data reporting by the paper's authors.

However, there is statistical literature on the multiplicity issue that is not followed well in the health sciences journals, including those from surgery. Readers are directed references [8, 25–29] for more information. Again, we would like to emphasize that, if you are a surgical researcher, your collaborative biostatistical colleague can help with the selection of a suitable method and provide references to support what was used.

Resolution of Clinical Scenario

Recall that in the PICOT we created, the outcome was complications. However, in the paper, there are three ways of considering complications: their rate, mean per patient, and their severity. The first of these was the primary outcome of the trial, so might be the most important for the patient and discussions with the patient. Barberan-Garcia et al. [1] report using a Chi-square test and a P value of 0.001, which was statistically significant at the 0.05 level; although the authors did not report the value of the Chi-square statistic. Using the same data as the authors, we showed how to compute the Chi-square statistic and found the P value was indeed less than 0.001! A flaw in this thinking is that the authors thought they were doing an ITT analysis; however, they were incorrect because they excluded 19 patients who were randomized from their analysis. A subsequent sensitivity analysis was not totally convincing since one of the analyses suggested that the P value was around 0.1 and so it would not be statistically significant. Thus, the benefit of the prehabilitation intervention would be 14%, which was less than what was considered as the minimum important difference of 20% in the sample size justification by the authors of the paper. The appraised paper's authors [1] also show the results of a relative risk analysis that was reported with a 95% confidence interval as their test statistic method. However, the numerical value of this test statistic was not reported. The paper's authors [1] report

stated that the RR was 0.5, 95% CI: 0.3 to 0.8, $P = 0.001$. We were able to use the same data as the paper's authors and our calculation gave 0.495, 95% CI: 0.325 to 0.755, which when rounded to 1 decimal place gave the same results as the authors [1]; however, with a P value of less than 0.001, just like we obtained with the Chi-square test. Note, both of these CIs excluded 1(one), which is the value assumed in the null hypothesis. We were able to redo the sensitivity analysis with the 19 patients that were left out and came to the same conclusion. However, one of the four extremes computed had a confidence interval that included 1 (one) from the null hypothesis. This provides doubt about the veracity of the data shown by the authors, suggesting that there may be no benefit to the patients.

For the mean complication per patient analysis, we were unable to do a sensitivity analysis, because we were not given any data on the 19 patients excluded, nor did we have their baseline characteristics to do a multiple imputation analysis. We were able to redo the analyses of the means found in the paper. The analysis of the means reported in the Barberan-Garcia et al. [1] paper was very similar to our computations.

For the severity of the complications, the authors report no data (Table 27.1) to support the sentence that they claimed there was no difference in the severity of complications. This is not quite credible as the large increases are seen in physical performance in the intervention group as well as the much shorter ICU length of stay. We do not find this conclusion of the authors credible.

Finally, the fact is that there are three different ways of expressing the complications; they would be derived from the same patients and so would be highly correlated measures of outcome, yet there was no mention of how these three analyses would reflect on the conclusions of this study. Certainly individual P values would not lead to the same conclusion for all three variables, even though that is what was reported, apart from severity. Here is where the multiplicity of testing issue should be discussed by Barberan-Garcia et al. [1] and how they proposed to handle it statistically, and if that had an

important impact on their conclusion for the outcome of complications.

Our overall conclusion is that the PICOT question could not be answered by the results of the appraised study [1]. Better evidence should be found that does not have the same methodological flaws as the paper we found seems to have. Maybe a trial older than 2018 would have used ITT properly and the complications would not have been coded in three different ways to confuse the reader.

Reporting Statistical Findings in the Surgical Literature

When summarizing the information that was reported in Table 27.1, many ideas were written about statistical issues that displayed a lack of reporting clarity, and we often disagreed with what Barberan-Garcia et al. [1] had said. The most serious issue from our viewpoint is the claim that an "Intention-To-Treat" analysis was done. This was NOT correct because the authors [1] chose to ignore the 19 patients that were randomized to the prehabilitation arm or not and reported on a total sample of 125 patients instead of the 144 randomized. We also could not identify from the author list in the appraised paper [1], a biostatistical collaborator. To rectify these concerns in the future, we suggest that surgical researchers always have a highly trained statistical collaborator as part of their research team to help write about the statistical issues more cogently. Readers interested in statistical reporting guidelines may be interested in references [25, 30–34].

Appendices

Appendix A: Articles Identified in the Literature Search

1. Skořepa P, Ford KL, Alsuwaylihi A, O'Connor D, Prado CM, Gomez D, Lobo DN. The impact of prehabilitation on outcomes in frail and high-risk patients undergoing major abdominal surgery: A systematic review and meta-analysis. Clin Nutr.

- 2024;43(3):629–648. <https://doi.org/10.1016/j.clnu.2024.01.020>.
2. Franssen RFW, Bongers BC, Vogelaar FJ, Janssen-Heijnen MLG. Feasibility of a tele-prehabilitation program in high-risk patients with colon or rectal cancer undergoing elective surgery: A feasibility study. *Perioper Med (Lond)*. 2022;11(1):28. <https://doi.org/10.1186/s13741-022-00260-5>.
3. Laurent H, Gravier FE. Preparing for thoraco-abdominal surgery: What's in it for the physiotherapist? *Rev. Mal Respir*. 2022;39(4):376–385. <https://doi.org/10.1016/j.rmr.2022.02.057>.
4. A van Beijsterveld C, Bongers BC, den Dulk M, Dejong CH, van Meeteren NL. Personalized community-based prehabilitation for a high-risk surgical patient opting for pylorus-preserving pancreaticoduodenectomy: A case report. *Physiother Theory Pract*. 2021;37(12):1497–1509. <https://doi.org/10.1080/09593985.2019.1709233>.
5. Fenton C, Tan AR, Abaraogu UO, McCaslin JE. Prehabilitation exercise therapy before elective abdominal aortic aneurysm repair. *Cochrane Database Syst Rev*. 2021;7(7):CD013662. <https://doi.org/10.1002/14651858.CD013662.pub2>.
6. Bongers BC, Dejong CHC, den Dulk M. Enhanced recovery after surgery programmes in older patients undergoing hepatopancreatobiliary surgery: What benefits might prehabilitation have? *Eur J Surg Oncol*. 2021;47(3 Pt A):551–559. <https://doi.org/10.1016/j.ejso.2020.03.211>.
7. Daniels SL, Lee MJ, George J, Kerr K, Moug S, Wilson TR, Brown SR, Wyld L. Prehabilitation in elective abdominal cancer surgery in older patients: Systematic review and meta-analysis. *BJS Open*. 2020;4(6):1022–41. <https://doi.org/10.1002/bjs5.50347>.
8. Thomas G, Tahir MR, Bongers BC, Kallen VL, Slooter GD, van Meeteren NL. Prehabilitation before major intra-abdominal cancer surgery: A systematic review of randomised controlled trials. *Eur J Anaesthesiol*. 2019;36(12):933–945. <https://doi.org/10.1097/EJA.0000000000001030>.
9. Wobith M, Acikgöz A, Grosser K, Weimann A. Preoperative cognitive function in very old patients: Influence on the complication rate and length of hospitalization. *Chirurg*. 2019;90(11):930–935. <https://doi.org/10.1007/s00104-019-01028-2>.
10. Barberan-Garcia A, Ubre M, Pascual-Argente N, Risco R, Faner J, Balust J, Lacy AM, Puig-Junoy J, Roca J, Martinez-Palli G. Post-discharge impact and cost-consequence analysis of prehabilitation in high-risk patients undergoing major abdominal surgery: Secondary results from a randomised controlled trial. *Br J Anaesth*. 2019;123(4):450–456. <https://doi.org/10.1016/j.bja.2019.05.032>.
11. Ausania F, Senra P, Meléndez R, Caballeiro R, Ouviaña R, Casal-Núñez E. Prehabilitation in patients undergoing pancreaticoduodenectomy: A randomized controlled trial. *Rev. Esp Enferm Dig*. 2019;111(8):603–608. <https://doi.org/10.17235/reed.2019.6182/2019>.
12. Barberan-Garcia A, Ubré M, Roca J, Lacy AM, Burgos F, Risco R, Momblán D, Balust J, Blanco I, Martínez-Pallí G. Personalised prehabilitation in high-risk patients undergoing elective major abdominal surgery: A randomized blinded controlled trial. *Ann Surg*. 2018;267(1):50–56. <https://doi.org/10.1097/SLA.0000000000002293>.
13. Debes C, Aissou M, Beaussier M. Prehabilitation. Preparing patients for surgery to improve functional recovery and reduce postoperative morbidity. *Ann Fr Anesth Reanim*. 2014;33(1):33–40. <https://doi.org/10.1016/j.annfar.2013.12.012>.

Appendix B: Sensitivity Analysis for Table 27.3 Using Chi-square Test and Minitab 18

- (a) Assume outcomes for missing patients were
all NO, for both arms

Comps\Arm	I,%	C,%	Total,%	Δ	Pearson X^2 , LR X^2	DF	P Values	Report P2
Yes	19,26	39,55	58,40	29	12.499, 12.793	1	0.000, 0.000	< 0.001, < 0.001
No	54	32	86					
Total	73	71	144					

Comps = Complications, Δ = %C – %I, LR = Likelihood Ratio, X^2 = Computed Chi-square statistic, DF = Degrees of Freedom, P2 = 2 tailed P value; Also for *b*, *c* and *d*

- (b) Assume outcomes for missing patients were
all NO for I and Yes for C

Comps\Arm	I,%	C,%	Total,%	Δ	Pearson X^2 , LR X^2	DF	P Values	Report P2
Yes	19,26	47,66	66,46	40	23.394, 24.077	1	0.000, 0.000	< 0.001, < 0.001
No	54	24	78					
Total	73	71	144					

- (c) Assume outcomes for missing patients were
all NO for C and Yes for I

Comps\Arm	I,%	C,%	Total,%	Δ	Pearson X^2 , LR X^2	DF	P Values	Report P2
Yes	30,41	39,55	69,48	14	2.760, 2.769	1	0.097, 0.096	No change
No	54	32	86					
Total	73	71	144					

- (d) Assume outcomes for missing patients were
all Yes, for both arms

Comps\Arm	I,%	C,%	Total,%	Δ	Pearson X^2 , LR X^2	DF	P Values	Report P2
Yes	30,41	47,66	77,53	25	9.115, 9.219	1	0.003, 0.002	No change
No	43	24	67					
Total	73	71	144					

- (e) Sensitivity analysis summary of all four
tables above: Think checker board! Using a
3 × 3 board

	I No		I Yes
C no	< 0.001 , < 0.001 , <i>a</i>	–	0.097, 0.096, <i>c</i>
	–	–	–
C yes	< 0.001 , < 0.001 , <i>b</i>	–	0.003 , 0.002 , <i>d</i>

Entries are P2 values, letters indicate the Tables above; Dashes (–) indicate parts of the sensitivity that we did not compute. Notice that the bolded cells, *a*, *b*, and *d* are all less than $\alpha = 0.05$, so are statistically significant; while *c* cell is not bolded so is not statistically significant

Appendix C: Risk of Bias Scoring of the Barberan-Garcia et al. [1] paper

Risk of Bias

A	Was the method of randomization adequate?	Yes
B	Was the treatment allocation concealed?	No
	Was knowledge of the allocated interventions adequately prevented during the study?	
C	1. <i>Was the patient blinded to the intervention?</i>	No
D	2. <i>Was the care provider blinded to the intervention?</i>	No
E	3. <i>Was the outcome assessor blinded to the intervention?</i>	Unsure
	Were incomplete outcome data adequately addressed?	
F	1. <i>Was the dropout rate described and acceptable?</i>	No, No
G	2. <i>Were all randomized participants analyzed in the group to which they were allocated?</i>	No
H	Are reports of the study free of suggestion of selective outcome reporting?	Unsure
	Other sources of potential bias:	
I	1. <i>Were the groups similar at baseline regarding the most important prognostic indicators?</i>	Unsure
J	2. <i>Were co-interventions avoided or similar?</i>	Unsure
K	3. <i>Was the compliance acceptable in all groups?</i>	Unsure
L	4. <i>Was the timing of the outcome assessment similar in all groups?</i>	Unsure
M	Is there a serious and fatal flaw with this study? (focus on the impact of selection bias, information bias, reporting errors, and confounding)	Flawed
	Are other sources of potential bias unlikely? Funding bias, conflict of interest statement Outcome measure not valid	Unsure

The Cochrane Risk of Bias Tool is available from Higgins et al. [3]

References

1. Barberan-Garcia A, Ubré M, Roca J, Lacy AM, Burgos F, Risco R, Momblan D, Balust J, Blanco I, Martinez-Palli. Personalised prehabilitation in high-risk patients undergoing elective major abdominal surgery: a randomized blinded controlled trial. *Ann Surg*. 2018;267(1):50–6.
2. Cadeddu MO, Farrokhyar F, Levis C, Haines AT, Thoma A for the Evidence-Based Surgery Working Group. Users' guide to the surgical literature; understanding confidence intervals. *Can J Surg*. 2012;55(3):207–11.
3. Higgins JPT, Altman DG, Gøtzsche, Jüni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JA, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. 2011;343:d5928.
4. Schünemann HJ, Guyatt GH. Commentary—good-bye M(C)ID! Hello MID, where do you come from? *Health Serv Res*. 2005;40(2):593–7.
5. Minitab Inc. Minitab [Internet]. [cited 2024 Apr 18]. Available from <http://www.minitab.com/en-us/>
6. van Buuren S. Flexible imputation of missing data. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2018.
7. Goodman SN. *p* Values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Amer J Epidemiol*. 1993;137(5):485–96.
8. Fraser DAS, Reid N. Crisis in science? Or crisis in statistics! Mixed messages in statistics with impact on science. *J Statist Res*. 2016;48-50(1):1–9.
9. Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: content, process, and purpose. *Am Stat*. 2016;70(2):129–33.
10. Mark DB, Lee KL, Harrell FE Jr. Understanding the role of P values and hypothesis tests in clinical research. *JAMA Cardiol*. 2016;1(9):1048–54.
11. Ioannidis JPA. The proposal to lower P values thresholds to .005. *JAMA*. 2018;319(14):1429–30.
12. Wild CJ, Pfannkuch M, Regan M. Towards more accessible conceptions of statistical inference. *J Roy Statist Soc A*. 2011;174(Part 2):1–23.
13. Hurlbert SH, Lombardi CM. Lopsided reasoning on lopsided tests and multiple comparisons. *Aust N Z J Stat*. 2012;54(1):23–42.
14. Altman DG, Machin D, Bryant TN, Gardner MJ. Statistics with confidence. 2nd ed. BMJ Books; 2011.
15. Streiner DL. One-tailed and two-tailed tests. *Statistics Commentary Series; Commentary #12. J Clin Psychopharmacol*. 2015;35(6):628–9.

16. CRAN (Comprehensive R Archive Network). The Comprehensive R Archive Network [Internet]. [cited 2024 Nov 29]. Available from: <http://CRAN.R-project.org>
17. Wheatley M. 5 key considerations when choosing open source statistical software. Data Informed. Big Data and Analytics in the Enterprise; n.d.. Available from: <https://medium.com/@limor.maayan/going-open-source-use-these-5-key-considerations-for-choosing-your-open-source-tool-92f358bf72f5>
18. Edwards H. A review of probability and statistics apps for mobile devices. In: Maker K, de Sousa B, Gould R, editors. Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics. (ICOTS July, 2014), Flagstaff AZ, Voorburg; 2014. p. 1–4.
19. McCullough BD, Yalata AT. Spreadsheets in the cloud—not ready yet. *J Stat Softw.* 2013;52(7) Available from <https://www.jstatsoft.org/article/view/v052i07/v52i07.pdf>
20. Melard G. On the accuracy of statistical procedures in Microsoft Excel 2010. *Comput Stat.* 2014;29(5):1095–128.
21. The R Foundation. The R Project for Statistical Computing [Internet]. [cited 2025 Apr 18] Available from: <https://www.r-project.org>
22. Pearson RK. Exploratory data analysis using R. Boca Raton: Chapman & Hall/CRC; 2018.
23. Thieme N. R Generation: The story of a statistical programming language that became a subcultural phenomenon. *Significance.* 2018;15(4):14–9.
24. Little RJ, D’Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs M, Murphy SA, Neaton JD, Rotnitzky A, Scharfstein D, Shih WJ, Siegel JP, Stern H. The prevention and treatment of missing data in clinical trials. *NEJM.* 2012;367(14):1355–60.
25. Mayo-Wilson E, Fusco N, Li T, Hong H, Canner JK, Dickersin K for the MUDS investigators. Multiple outcomes and analyses in clinical trials create challenges for interpretation and research synthesis. *J Clin Epidemiol.* 2017;86:39–50.
26. Perneger TV. What’s wrong with Bonferroni adjustments? *BMJ.* 1998;316(7137):1236–8.
27. Cook RJ, Farewell VT. Multiplicity considerations in the design and analysis of clinical trials. *J Roy Statist Soc A.* 1996;159(1):93–110.
28. US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Multiple Endpoints in Clinical Trials [Internet]. [cited 2024 Apr 18]. Available from <https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm536750.pdf>
29. Bland JM, Altman DG. Multiple significance tests; the Bonferroni method. *Statistics notes.* *BMJ.* 1995;310(6973):170.
30. Livingston EH. Introducing the JAMA guide to statistics and methods. *NEJM.* 2014;312(1):35.
31. Aslan D, Sandberg S. Simple statistics in diagnostic tests. *J Med Biochem.* 2007;26(4):309–13.
32. Lydersen S. Statistical review: frequently given comments. *Ann Rheum Dis.* 2015;74:323–5.
33. Lang TA, Altman DG. In: Smart P, Maisonneuve H, Polderman A, editors. Basic Statistical Reporting for Articles Published in Biomedical Journals: The “Statistical Analyses and Methods in the Published Literature” or “The SAMPL Guidelines”. *Science Editors’ Handbook.* European Association of Science Editors; 2013. p. 1–8.
34. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviations from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol.* 2014;14:135.

Confidence Intervals

28

Lily J. Park, Jessica Bogach, Lawrence Mbuagbaw,
and Margherita Cadeddu

Abbreviations

CI	Confidence Interval
GRADE	Grading of Recommendations, Assessment, Development, and Evaluations
RCT	Randomized Controlled Trial
MCID/MID	Minimal Clinical Important Difference/Minimal Important Difference
MBP	Mechanical Bowel Preparation
oAB	oral Antibiotics
RR	Relative Risk
PICOT	Population, Intervention, Comparison, Outcome, Timing
SSI	Surgical Site Infection
SD	Standard Deviation
SE	Standard Error
ARR	Absolute risk reduction
FI	Fragility Index
CRM	Circumferential resection margin
SR	Systematic Review

L. J. Park · J. Bogach
Division of General Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
Department of Health Research Methods,
Evidence, and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: lily.park@medportal.ca;
bogachj@mcmaster.ca

Introduction

Confidence intervals (CIs) are essential for researchers to communicate their study results and for clinicians to interpret those results for application in their clinical practice. With the ongoing debate regarding appropriate use and interpretation of p-values, there is a greater emphasis on the importance of understanding CIs [1, 2].

Beyond simply determining statistical significance, CIs provide information on the precision of an estimated population parameter. By determining the interval of where an estimate ranges, a CI allows us to quantify the level of uncertainty in the inferences we make about the larger population based on findings from our sample data [3]. Interpreting the range of confidence with predefined error probabilities helps us to shift our focus from statistical significance to practical relevance of the observed effects.

L. Mbuagbaw
Department of Health Research Methods,
Evidence, and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: mbuagblc@mcmaster.ca

M. Cadeddu (✉)
Division of General Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: mcadeddu@stjosham.on.ca

CI is usually expressed as a percentage that helps determine the level of confidence that the true population parameter lies within the calculated CI [3]. For example, a 95% CI represents a 95% chance of constructing an interval that will cover the true population parameter of interest, before the sample is drawn [3]. This means that if the same population were to be sampled many times and a CI were to be constructed for each experiment, the resulting intervals would contain the true population parameter 95% of the time. The confidence interval is a measure of precision and reflects how confident we are that the true population value lies within the range of estimates. Therefore, if the results of an experiment are associated with a narrow CI at a selected confidence level (e.g., 95%) this would suggest greater precision and you would have more confidence in your results. However, if at the selected confidence level, the results are associated with a wide CI, this would suggest imprecision, and you would have less confidence in your results.

Clinical Scenario

You are a General Surgeon arranging an elective laparoscopic low anterior resection for a patient that was assessed to have resectable high rectal cancer. At the patient's preoperative appointment, you discuss that they will be given mechanical bowel preparation (MBP) as well as oral antibiotics. The patient states that they want to avoid any "unnecessary medications" and asks if the antibiotics are entirely necessary. You recall there have been recent changes in recommendations for preoperative bowel preparation and decide to perform an updated literature review.

You develop a clinical research question (see Chap. 3) and frame it in the following PICOT (Population, Intervention, Comparison, Outcome, Time period) format:

P: In patients undergoing elective low anterior resection for rectal cancer
 I: Does the use of MBP and oral antibiotics (oAB)
 C: compared to MBP alone

O: decrease the risk of surgical site infections (SSI) and/or mortality
 T: within 30 days from surgery

The Search

Using keywords, you perform a literature review as described in Chap. 4 and come across the following article: *Preoperative combined mechanical and oral antibiotic bowel preparation for preventing complications in elective colorectal surgery* by Willis et al. (Appendix 1) [4].

This is a Cochrane systematic review and meta-analysis that compared postoperative outcomes between patients that received mechanical bowel prep (MBP) and oral antibiotics (oAB) versus MBP alone. Considering the clinical question at hand, we will focus on the surgical site infection (SSI) and mortality outcomes only.

The review reported a relative risk (RR) of 0.56 with a 95% CI of 0.42–0.74 for the outcome of SSI. The RR for mortality was 0.87 (95% CI 0.27–2.82) [4]. You interpret this to mean that compared to MBP only, MBP + oAB is associated with lower rates of SSI and possibly lower mortality but would like to understand better what the corresponding confidence intervals mean.

Understanding Confidence Intervals

Within every research inquiry, the goal is to address a knowledge gap. One strategy in surgical research involves selecting a sample population for analysis in a study with the intention of extrapolating the findings to a broader patient population [3]. In our example, we are trying to determine if prophylactic MBP + oAB is superior to prophylactic MBP alone for elective colorectal procedures.

As a systematic review and meta-analysis, Willis et al. pooled data from up to 16 randomized trials involving 3917 patients to offer insight into our research question [4]. The combination of multiple trials and patients allows for a larger

sample size and, therefore, a more robust summary of findings regarding the relationships of interest. However, interpretation of meta-analysis results should be made with consideration of limitations stemming from potential sources of variation or bias within and between the trials that have been combined (Table 28.1). Also see Chap. 16 The Surgeon’s Guide to Systematic Review and Meta-Analysis.

For our scenario, we would like to determine the true risk of SSI and mortality with MBP + oAB compared to the risk of that with MBP alone across *all* patients undergoing elective colorectal surgery. Because we cannot know this true RR, we can use a collection of different trial populations, to garner the largest sample population we can and estimate the truth. Although this helps the external validity of the trial, it is still not inclusive of the entire population of interest. Therefore, we expect that the value we obtain is not identical to the true, unknown, population value, and the CI is a measure of imprecision and reflects how confident we are that the true population value lies within the range of estimates. CIs are essentially the expression of the variability and uncertainty of the study results. The imprecision of our estimate can come from multiple sources, including sampling error, small

sample size, inaccurate measurement, and the prespecified probability that we have selected (i.e., 99% vs. 95%) [5].

To understand how we get to CIs from raw study data, an understanding of the standard deviation (SD) and standard error (SE) of the mean is necessary. For a certain characteristic that is being studied in a random sample of a population, there will be variations in the results. For example, if wound infection rates are being measured in 10% of the hospital’s surgical population, and each month a random sample of patients is measured, this estimate will vary each month around a central value. This creates a distribution of values. From each month, a mean wound infection rate can be calculated along with a standard deviation (SD) which shows, on average, how much the estimates varied from the mean infection rate. Because we are using a sample, we use the SE as a measure of how far off we are from the true population mean [6]. The SE is calculated by dividing the SD by the square root of the sample size. Therefore, the larger the sample size, the more it approximates the population and the smaller our SE will be. For hypothesis testing and estimation, we use the SE: it is a key component in CI calculation.

Table 28.1 Sources of systematic errors and bias impacting meta-analyses

Possible systematic errors and bias [3]	Example
Systematic differences between the intervention group and the control group	Since all included trials in the Willis et al. review were randomized, there should not be systematic differences between the two groups, if randomization was done correctly in all trials. However, this should be of concern if a systematic review and meta-analysis includes comparative observational studies that have not been randomized
Differences in patient characteristics between trials	Differences in the distribution of rectal versus colonic cancer resections between trials; differences in trial inclusion and exclusion criteria between trials; differences in patient characteristics (age, sex, disease severity, etc.)
Different administration of the intervention	Differences in specific oral antibiotic regimen (coverage, dosage, frequency) between trials
Differences in outcome definition	Definition of anastomotic leak may vary between trials (clinical and/or radiologic)
Differences in blinding of group allocation	In trials that are not blinded, knowledge of the treatment they are receiving may affect patient or provider behavior and lead to biased assessment of the effectiveness of the intervention
Loss to follow-up/adherence to treatment	Unbalanced participant compliance and retention in the trial may lead to bias if it is associated with the treatment received

Table 28.2 Factors that change the width of your confidence interval

Narrows CI	Widens CI
Large sample size	Small sample size
Smaller standard error	Larger standard error
Low degree of confidence (i.e., 90%)	High degree of confidence (i.e., 99%)
Homogeneous data	Large amount of variability

How Is the Confidence Interval Calculated?

The method of calculating the CI depends on the type of estimate you are creating the interval for: is it a point estimate or a ratio? If the interval is for a mean, compared to an odds ratio, the calculation is done differently. However, each calculation takes into account two important factors: the degree of confidence (i.e., 95%) and the SE (Table 28.2) [14].

The calculation for CI is as follows:

- $CI = (Parameter - Z \text{ value} \times SE, Parameter + Z \text{ value} \times SE)$

Degree of confidence—we chose a prespecified Z value depending on our confidence level

- 90% smaller confidence level
- 99% larger confidence level

SE: This helps determine the precision of the sample estimate compared to the population parameter. The SE calculations depend on:

- Number of events
- Sample size
- Variability of data
- Larger Standard Errors lead to larger CI

The larger interval will be more likely to contain the true population value but will also be more likely to cross into “no difference in effect” territory.

Hypothesis Testing

Similar to p-values, we can use the 95% CI to test a null hypothesis [7]. The null hypothesis is a fundamental statistical concept that states there is no significant difference between groups and that any observed differences are due to chance. When we test the null hypothesis, we are trying to surmount evidence to either reject or fail to reject the null hypothesis, that is, that there is no difference between the groups of interest.

It is important to note that the null hypothesis value is chosen based on the specific question you are trying to address and the context of your study. For instance, if you were testing the difference in proportions or means, the null value would be 0. This is because there is zero or no difference in the proportions or the mean value of certain outcomes between two groups. However, if you are testing the relative risk or odds ratio, the null value would be 1 because a ratio of 1 between two groups suggest there is no difference between them.

In the systematic review used for our scenario, the null hypothesis would be that the risk of SSI and mortality are the same with and without the addition of oAB to the preoperative MBP. Therefore, the relative risk for this would be 1. So, if our 95% CI crosses our null value, we cannot reject our null hypothesis and there is no difference in SSI or mortality with or without oAB. In our example, oAB reduces the risk of

SSI by 44% with a 95% CI of 0.42 to 0.74, suggesting that this reduction can be as large as 58% or as small as 26%. This shows that the point estimate and the CIs include only benefits. However, the result for mortality was a RR of 0.87 with a 95% CI of 0.27 to 2.82. In other words, the addition of oAB can reduce the risk of death by 13% with reduction possibly being as large as 73%. However, oAB could also increase the risk of mortality by 182%. Although the point estimate suggests benefit, the 95% CI crosses the null value and includes the possibility of increased risk. Thus, we cannot conclude that there is a difference in mortality with or without oAB.

We also must consider to what we are comparing our estimate. For example, if we are taking a single sample and comparing it against a known value (i.e., is the mean of our sample greater than the expected population mean?), our only source of variability is our estimation from the sample. If the mean population value is

included within the CI, we can say that there is no statistically significant difference between our sample and the population [9]. However, when comparing two groups, we are introducing more sources of variability. Often when reading literature, when faced with overlapping CIs, clinicians instinctively think that the two populations are not significantly different. However, when the two values are formally tested (e.g., with a t-test), it is possible with overlapping CIs that they will show a statistically significant difference [9]. In fact, the CI's can overlap up to 29% and still have significant p-values [9]. Therefore, we must be careful in our interpretation of CIs that are overlapping when comparing two estimated values (Table 28.3). When comparing groups, it is more useful to create an estimate and CI for the relationship between the two results (i.e., mean difference or relative risk) rather than comparing the CIs for each result, as this will be less susceptible to misinterpretation [4].

Table 28.3 Examples of 95% confidence intervals and their appropriate interpretation

Study	Type of result	Null value	Result	Reported confidence interval (CI)	Interpretation
Coakley et al. Prophylactic Ureteral Catheters for Colectomy: A National Surgical Quality Improvement Program-Based Analysis [7]	Odds ratio	1	0.45	95% CI 0.25–0.81	The use of prophylactic ureteric catheter placement decreases the risk of ureteric injury
Columbo et al. A Meta-analysis of the Impact of Aspirin, Clopidogrel, and Dual Antiplatelet Therapy on Bleeding Complications in Noncardiac Surgery [8]	Relative risk	1	0.96	95% CI 0.76–1.22	There is no difference in bleeding rates requiring re-intervention with the use of aspirin
Gianotti et al. Preoperative Carbohydrate Load Versus Placebo in Major Elective Abdominal Surgery (PROCY) [9]	Relative difference	0	0.003	95% CI –0.053 to 0.059	The use of oral carbohydrate load does not decrease the risk of wound infection
Koullouros et al. The role of oral antibiotic prophylaxis in prevention of surgical site infection in colorectal surgery [10] (Two trials from the meta-analysis) [11, 12].	Odds ratio	1	1.27	Espin-Basany et al. [12] 95% CI 0.48–3.38	In the first trial there is no difference in surgical infection rate with the use of oral antibiotic prophylaxis. In the second study there is a reduction in surgical infection rate. However, we cannot infer any difference between the two trials because of the overlapping confidence intervals without a formal statistical test
			0.30	Sadahiro et al. [11] 95% CI 0.11–0.79	

Minimally Important Difference (MID) and GRADE

In addition to statistically significant differences, we must also consider the Minimally Important Difference (MID) when evaluating evidence to apply to our practice. MID is a concept that represents the smallest change that must occur in a clinical outcome to be considered meaningful to patients and healthcare professionals [13–15]. It helps to contextualize the magnitude of treatment effects and informs decisions regarding clinical relevance of interventions. This can be specified overtly, or clinicians can determine this independently [14].

MID plays a large role within the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework, which is widely used to evaluate the quality of evidence and strength of recommendations in systematic reviews and practice guidelines [16, 17]. Understanding CIs and MID in the context of GRADE is critical to the clinician’s assessment of applicability of review findings to their practice.

Firstly, we should note what the MID is, how it was determined by the authors, and whether it was determined a priori. The protocol for the review by Willis et al. stated a priori that for SSI, an absolute risk reduction (ARR) of 5% would be considered clinically relevant [18]. Based on the

formula of Relative Risk Reduction = $100\% \times [1 - \text{Risk Reduction}]$, the authors deduced that an RR of less than or equal to 0.95 would achieve the desired ARR of 5%. The authors do not provide a justification for how they came to decide on the 5%—this may not always be provided. Therefore, the reader/clinician must evaluate and decide whether they agree that this predetermined difference should be considered the MID.

Specifically, CI and MID impacts the precision domain of the GRADE assessment framework [19]. Precision describes the degree of confidence in the estimated pooled effect. Simply, it is an indicator of the narrowness of the CI around the effect estimate. A narrower CI indicates higher precision, while a wider interval suggests lower precision. A narrower CI is also more likely to provide meaningful results. For instance, Fig. 28.1a represents the pooled effect estimate and the corresponding 95% CI for the outcome of SSI. The vertical line $X = 1$ represents the null hypothesis, that is, there is no statistically significant difference in SSI events between MBP + oAB and MBP only. The orange dotted line in Fig. 28.1a represents the a priori MID established by Willis et al. of RR 0.95. The systematic review by Willis et al. reported RR 0.56 with a 95% CI of 0.42–0.74 for the incidence of SSI in those who received MBP + oAB versus those who received MBP alone. The large

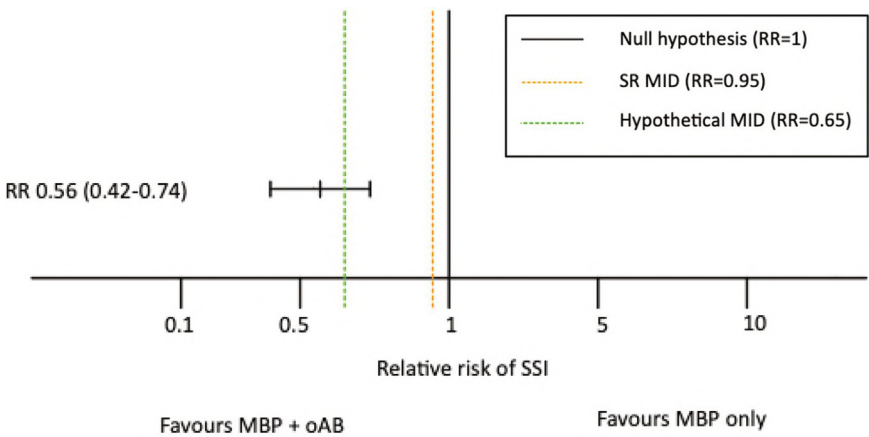


Fig. 28.1a Results of SSI comparing MBP + oAB versus MBP alone, where a RR of 0.95 are compared to a theoretical RR of 0.65 as the minimally important difference (MID)

sample size (3917) and number of events (410) likely contributed to the narrow CI, which does not cross the threshold of no statistical difference (i.e., $x = 1$) nor the threshold of no clinically important difference (i.e., $X = 0.95$). This gives us more confidence in the benefit of MBP + oAB over MBP only with respect to SSI events and therefore, the GRADE assessment was not downgraded for precision here. Had the MID been larger and required a greater relative risk reduction of 0.65 (vertical green line), the results would have been considered statistically significant but not clinically important since the CI would cross $X = 0.65$ but not $X = 1$. In this case, the GRADE assessment would have likely been downgraded by 1 for precision [16, 19]. Altogether, the results demonstrate clinical and statistical differences between MPB + oAB and MBP only, which are important to clinicians if

we are to inform changes in our practice. Fig. 28.1b demonstrates the effect estimate and 95% CI for mortality. Although MID was not determined a priori for mortality, this is a common practice as any difference in mortality is considered important to patients and clinicians.

Other hypothetical scenarios that can result with respect to a predetermined MID are demonstrated in Fig. 28.2. Again, if the green represents the MID, and $X = 1$ represents the null value, we can interpret the individual results as follows: A would be considered statistically significant favoring double therapy (MBP + oAB) but not clinically important since confidence interval does not contain $X = 1$ but also does not cross the MID value. B would be considered not statistically significant nor clinically important; in this scenario, the null hypothesis cannot be rejected because the CI crosses $X = 1$. C would

Fig. 28.1b Results of mortality comparing MBP + oAB versus MBP alone, where the minimally important difference (MID) was not determined a priori

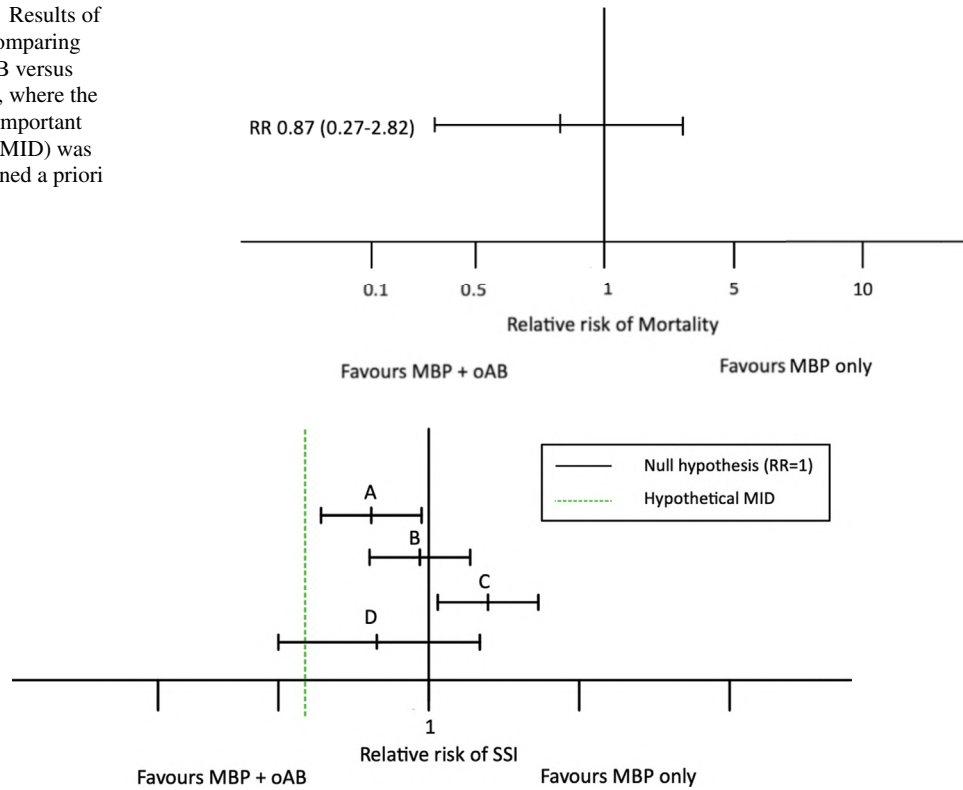


Fig. 28.2 Hypothetical results comparing patients given MBP + oAB versus MBP alone, where the dotted green line represents the minimally important difference (MID). (a) Interval is statistically significant favoring dual therapy but

not clinically important. (b) Interval is neither statistically significant nor clinically important. (c) Interval is statistically significant favoring monotherapy. (d) Interval is neither statistically significant nor clinically important

be considered statistically significant in favoring monotherapy with MBP only. D would be considered neither statistically significant nor clinically important considering the CI crosses both the null and MID values. In the review by Willis et al. mortality resembled scenario B where there were no statistically significant difference (RR 0.87 95% CI 0.27–2.82) [4]. The small sample size (639) and small number of events (11) likely contributed to the large CI. This wide interval encompasses the possibility of considerable benefit (i.e., relative risk of death with MBP + oAB is 0.27 times the risk of death with MBP only) as well as the possibility of considerable harm (i.e., relative risk of death is 2.82 times higher in MBP+oAB versus MBP alone) [4]. Due to this imprecision, the GRADE assessment was justifiably downgraded by 1 for the mortality outcome.

Why Are CIs Different from P-Values?

Biomedical literature has a generally accepted level of significance, with an alpha of 0.05. Although this is convention, it is somewhat arbitrary [11]. Using a p-value alone, we tend to dichotomize outcomes as significant or nonsignificant, even if p values are very similar. For example, with a p-value of 0.049 we would reject a null hypothesis, but with a p-value of 0.051, we would not reject it [20]. Walsh, et al. showed that in major randomized controlled trials published in high impact journals, it is often a single event that can push the p-value into statistical significance [21]. Additionally, even with a statistically significant result, there is no information about magnitude of difference and clinical relevance [13].

The fragility index (FI) is a metric that describes the minimum number of participants in a trial whose status would have to change from a nonevent to an event to turn a statistically significant result to a nonsignificant result [21, 22]. A lower FI describes a more fragile outcome. It has become increasingly recognized that many randomized trials that guide treatment decisions

have low FI and may not be robust. Thus, interpretation of a study through p-values alone can push clinicians toward clinical decisions that may not be strongly supported.

CIs do suffer from the same fragility, in that if the CI covers our “null value” we would reject a statistically significant difference. This range can also be influenced by single data points. However, by using CIs, we are informed about the direction and magnitude of the effect.

Why Use CIs?

Cumming and Finch (2001) propose four reasons why CIs should be reported [23]:

1. By providing point and interval information, it facilitates understanding and interpretation.
2. CIs allow us to perform null hypothesis testing, but by looking at the range of values we can look at other hypotheses as well, not just the null hypothesis.
3. CIs are very useful in meta-analysis and evidence synthesis where we are comparing and combining different data.
4. They provide information about precision.

Back to the Scenario

Based on your understanding of CIs, you review the article by Willis et al. This was a Cochrane systematic review and meta-analysis looking at the difference in MBP + oAB versus MBP alone in various morbidity and mortality outcomes [4]. It reports a RR of 0.56 (95% CI 0.42–0.74) for SSI and RR 0.87 (95% CI 0.27–2.82) for mortality [4]. You ask yourself whether you can apply this article to your patient (Table 28.4).

Since the results apply to your patient and show a significant decrease in SSI with no difference in mortality based on moderate certainty of evidence, you discuss this with your patient and advocate for the addition of oAB to the MBP in preparation for their surgery. Although there is a greater risk of experiencing nausea/vomiting with the oAB you discuss that if the patient were

Table 28.4 Questions about article validity and applicability

Question	Answer
<i>Is this study valid?</i>	Systematic review and meta-analysis of randomized controlled trials that were considered to have low to some concerns of bias. Overall large number of participants. For our outcomes of interest, overall certainty of evidence was moderate SSI and moderate mortality
<i>What are the results?</i>	Treatment effect: We interpret the RR to suggest that MBP + OAB decreases the risk of SSI where the CI does not include 1, and is above a previously determined MID. We interpret this as a statistically significant result. We interpret the mortality RR to suggest that there is no difference in the risk of death between MBP + oAB and oAB alone
<i>Can I apply them to my practice?</i>	Is the population similar to my patient? 17 of the 21 included studies (3917 participants) investigated MBP + OAB versus MBP alone. All study participants underwent elective colorectal surgery for benign and malignant indications. Therefore, the results are generalizable to your patient. Are the treatment benefits worth the potential harm and costs: greater experience of nausea/vomiting and abdominal pain with the addition of OAB (3 studies, 545 participants), no difference in c difficile-related diarrhea

to develop a postoperative SSI, this would warrant antibiotics anyway in addition to potentially more invasive interventions like drain insertions and reoperation.

Following your discussion, the patient feels they prefer adding oral antibiotics to their preoperative regimen.

They are now wondering whether they will require an open versus a laparoscopic surgery. Your training consisted mostly of the laparoscopic approach and so you discuss that you will begin this way with the possibility of converting to open, if needed. Following your encounter, you wonder how the shift in practice occurred from open to laparoscopic surgeries over the years.

You find that there are many non-inferiority trials as well as a systematic review and non-inferiority meta-analyses on the topic. To understand the shift in practice further, you delve into these studies and notice many use a non-inferiority design. You wonder how to interpret CIs in this context.

Non-inferiority Trials

Non-inferiority trials have become increasingly commonplace and are important to understand and interpret. Non-inferiority studies help establish that a novel treatment’s effectiveness is not

substantially less than the existing standard (See Chap. 13 on Non-Inferiority Randomized Controlled Trials) [24, 25]. Therefore, noninferiority designs are based on comparing confidence interval of treatment effects and predefined clinically acceptable (non-inferior) margins [25, 26].

There are many reasons why a researcher may choose non-inferiority trials over superiority trials. These include: ethical considerations, clinical relevance, limited resources, or a context in which demonstrating superiority may be challenging [24, 25]. The non-inferiority approach was popular among trialists investigating laparoscopic versus open resection for rectal surgeries due to the well-established short-term benefits of laparoscopic surgery at the time including decreased blood loss, earlier return to bowel movement, and shorter length of hospital stay [27, 28]. This meant that if non-inferiority were to be established for long-term oncologic outcomes, it would suffice in supporting the laparoscopic approach. However, the cost of non-inferiority designs lies in the challenge of understanding and interpreting the trial results as this task can be more difficult compared to superiority trials.

In the meta-analysis by Acuna et al. 14 RCTs were included [28]. The primary outcomes were measures of surgical resection quality including the number of positive circumferential resection

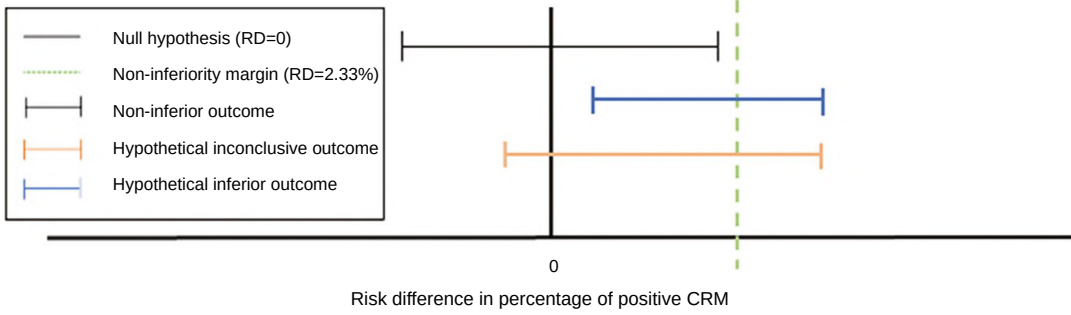


Fig. 28.3 Investigating non-inferiority of laparoscopic to open colorectal surgery in the outcome of positive circumferential resection margin (CRM), where the dotted green line represents the non-inferiority margin. Black: results

from the Acuna et al. review, demonstrating non-inferiority. Orange: hypothetical results demonstrating inconclusive outcome. Blue: hypothetical results demonstrating inferior outcome

margin (CRM) [28]. Much like MID margins, it is important to analyze the predetermined non-inferiority margins and their justifications. In this case, non-inferiority margins were determined through the mean margin generated by a Delphi consensus panel, which was a risk difference in the percentage of positive CRM of 2.33 [28]. The study results report that laparoscopic resection was non-inferior compared with open resection for the rate of positive CRM with a risk difference of 0.79%, 90% CI -0.46 to 2.04 [28]. Considering this confidence interval does not cross the margin of non-inferiority (2.33) and does cross the null value in risk difference of 0, we can conclude that laparoscopic surgery is non-inferior to open surgery with respect to CRM. Had the confidence interval crossed the predetermined non-inferiority margin, laparoscopic surgery would have been considered not non-inferior to open surgery (Fig. 28.3).

Understanding confidence intervals is pivotal for interpreting and drawing meaningful conclusions from non-inferiority trials. By considering the bounds of the CI in relation to a predefined non-inferiority margin, we can understand whether the effect of a new treatment lies within an acceptable and clinically relevant range. This nuanced approach mitigates the potential for misinterpretations that can occur when we rely solely on point estimates or p-values. In doing so, CIs can help support the robustness and credibility of clinical decisions made based on trial outcomes.

Key Points for Clinicians

- Determine if the study you are reviewing is valid.
- Does your CI include the null hypothesis value? Does it include your MID?
- CIs are expressions of the stability of the study results and show a range of results that likely contain the true population value with a certain degree (95%) of confidence.
- CIs are important to understand and report, as they inform us about our estimate and allow hypothesis testing.
- Care must be taken when interpreting CIs (overlapping CIs can still be significant).
- Intervals are impacted by our sample size, sample variability and the preselected degree of confidence (i.e., 95%).

Appendix 1

Search Strategy (Medline OVID, from January 1, 2020 to December 31, 2023)

1. Exp Colorectal Surgery/
2. (bowel adj2 prep*).mp.
3. Polyethylene Glycols/
4. (mechanical bowel prep*).mp.
5. Antibiotic Prophylaxis/
6. Anti-Bacterial Agents/
7. (oral adj2 antibiotic*).mp.

8. 2 or 3 or 4
9. 5 or 6 or 7
10. 1 and 8 and 9

References

1. Li G, Walter SD, Thabane L. Shifting the focus away from binary thinking of statistical significance and towards education for key stakeholders: revisiting the debate on whether it's time to de-emphasize or get rid of statistical significance. *J Clin Epidemiol*. 2021;137:104–12.
2. Tijssen JGP. More confidence intervals and fewer p values: a positive trend? *J Am Coll Cardiol*. 2021;77(12):1562–3.
3. Hazra A. Using the confidence interval confidently. *J Thorac Dis*. 2017;9(10):4125–30.
4. Willis MA, Toews I, Sohtau SL, Kalff JC, Meerpohl JJ, Vilz TO. Preoperative combined mechanical and oral antibiotic bowel preparation for preventing complications in elective colorectal surgery. *Cochrane Database Syst Rev* [Internet] 2023 [cited 2023 Jun 8];(2). Available from: <https://www-cochranelibrary-com.libaccess.lib.mcmaster.ca/cdsr/doi/10.1002/14651858.CD014909.pub2/full>.
5. Altman DG. Why we need confidence intervals. *World J Surg*. 2005;29(5):554–6.
6. Garg PK, Mohanty D. Mean (standard deviation) or mean (standard error of mean): time to ponder. *World J Surg*. 2013;37(4):932.
7. Coakley K, Kasten K, Sims S, Prasad T, Heniford B, Davis B. Prophylactic ureteral catheters for colectomy: a national surgical quality improvement program-based analysis. *Dis Colon Rectum* [Internet] 2018 Jan [cited 2023 Aug 9];61(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/29215477/>.
8. Columbo JA, Lambour AJ, Sundling RA, Chauhan NB, Bessen SY, Linshaw DL, Kang R, Riblet NBV, Goodney PP, Stone DH. A meta-analysis of the impact of aspirin, clopidogrel, and dual antiplatelet therapy on bleeding complications in noncardiac surgery. *Ann Surg*. 2018;267(1):1–10.
9. Gianotti L, Biffi R, Sandini M, Marrelli D, Vignali A, Caccialanza R, Viganò J, Sabbatini A, Di Mare G, Alessiani M, Antomarchi F, Valsecchi MG, Bernasconi DP. Preoperative oral carbohydrate load versus placebo in major elective abdominal surgery (PROCY): a randomized, placebo-controlled, multicenter, phase III trial. *Ann Surg*. 2018;267(4):623–30.
10. Koullouros M, Khan N, Aly EH. The role of oral antibiotics prophylaxis in prevention of surgical site infection in colorectal surgery. *Int J Color Dis*. 2017;32(1):1–18.
11. Sadahiro S, Suzuki T, Tanaka A, Okada K, Kamata H, Ozaki T, Koga Y. Comparison between oral antibiotics and probiotics as bowel preparation for elective colon cancer surgery to prevent infection: prospective randomized trial. *Surgery*. 2014;155(3):493–503.
12. Espin-Basany E, Sanchez-Garcia JL, Lopez-Cano M, Lozoya-Trujillo R, Medarde-Ferrer M, Armadans-Gil L, Alemany-Vilches L, Armengol-Carrasco M. Prospective, randomised study on antibiotic prophylaxis in colorectal surgery. Is it really necessary to use oral antibiotics? *Int J Color Dis*. 2005;20(6):542–6.
13. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*. 1986;292(6522):746–50.
14. Cook CE. Clinimetrics corner: the minimal clinically important change score (MCID): a necessary pretense. *J Man Manip Ther*. 2008;16(4):E82–3.
15. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA*. 2014;312(13):1342–3.
16. Schunemann H, Brozek J, Guyatt GH, Oxman AD. GRADE Handbook [Internet]. [cited 2023 Apr 6]. Available from: <https://training.cochrane.org/resource/grade-handbook>.
17. What is GRADE? | BMJ Best Practice [Internet] [cited 2023 Apr 4]. Available from: <https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/>.
18. Willis MA, Toews I, Meerpohl JJ, Vilz TO. Preoperative combined mechanical and oral antibiotic bowel preparation for preventing complications in elective colorectal surgery. *Cochrane Database Syst Rev*. 2022;2022(1):CD014909.
19. Zeng L, Brignardello-Petersen R, Hultcrantz M, Mustafa RA, Murad MH, Iorio A, Traversy G, Akl EA, Mayer M, Schünemann HJ, Guyatt GH. GRADE guidance 34: update on rating imprecision using a minimally contextualized approach. *J Clin Epidemiol*. 2022;150:216–24.
20. Piccirillo JF. Improving the quality of the reporting of research results. *JAMA Otolaryngol Head Neck Surg*. 2016;142(10):937–9.
21. Walsh M, Srinathan S, McAuley D, Mrkobrada M, Levine O, Ribic C, Molnar A, Dattani N, Burke A, Guyatt GH, Thabane L, Walter S, Pogue J, Devereaux PJ. The statistical significance of randomized controlled trial results is frequently fragile: a case for a fragility index. *J Clin Epidemiol* [Internet] 2014 Jun [cited 2023 Aug 9];67(6). Available from: <https://pubmed.ncbi.nlm.nih.gov/24508144/>.
22. The Fragility Index in Randomized Clinical Trials as a Means of Optimizing Patient Care | Trauma and Injury | JAMA Surgery | JAMA Network [Internet] [cited 2023 Aug 30]. Available from: <https://jamanetwork.com/journals/jamasurgery/article-abstract/2712857>.
23. Cumming G, Finch S. A primer on the understanding, use, and calculation of confidence intervals that are

- based on central and noncentral distributions. *Educ Psychol Meas.* 2001;61(4):532–74.
24. Snapinn SM. Noninferiority trials. *Curr Control Trials Cardiovasc Med.* 2000;1(1):19–21.
25. Hahn S. Understanding noninferiority trials. *Korean J Pediatr.* 2012;55(11):403–7.
26. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. *Trials.* 2011;12(1):106.
27. Schwenk W, Haase O, Neudecker J, Müller JM. Short term benefits for laparoscopic colorectal resection. *Cochrane Database Syst Rev.* 2005;2005(3):CD003145.
28. Acuna SA, Chesney TR, Ramjist JK, Shah PS, Kennedy ED, Baxter NN. Laparoscopic versus open resection for rectal cancer: a noninferiority meta-analysis of quality of surgical resection outcomes. *Ann Surg.* 2019;269(5):849–55.

Power and Sample Size

29

Eric Duku, Jessica Murphy, Emily Dunn,
Achilles Thoma, and Charles H. Goldsmith

Abbreviations

RCT	Randomized controlled trial
TAP	Transversus abdominis plane
MCID	Minimal clinically important difference
MID	Minimal important difference
NRS	Numerical Rating Scale
ES	Effect size
HRQOL	Health-related quality of life
VAS	Visual Analog Scale
BMI	Body Mass Index
PAR	Postanesthetic recovery
PODs	Postoperative days
POD1	Postoperative day 1
SD	Standard deviation

E. Duku (✉)
Department of Health Research Methods, Evidence,
and Impact, McMaster University,
Hamilton, ON, Canada

Offord Centre for Child Studies, Department
of Psychiatry and Behavioral Neurosciences,
McMaster University, Hamilton, ON, Canada
e-mail: duku@mcmaster.ca

J. Murphy
Division of Physical Medicine & Rehabilitation,
Department of Medicine, McMaster University,
Hamilton, ON, Canada
e-mail: murphj11@mcmaster.ca

E. Dunn
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: dunne4@mcmaster.ca

Introduction

A frequent criticism of published surgical studies pertains to whether a study has an adequate sample size. In other words, readers are hesitant that the chosen sample size would enable the trial to detect a clinically meaningful difference between surgical interventions, if one exists. This criticism applies to all study designs; however, it is particularly relevant to the randomized controlled trial (RCT), which is touted to produce the highest level of evidence for primary studies. While RCTs are the preferred study design, they can be incredibly difficult to implement in surgery. Many reasons have been given for the small number of RCTs performed in surgery, with recruitment difficulties being one of the most com-

A. Thoma
Division of Plastic Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: athoma@mcmaster.ca

C. H. Goldsmith
Department of Health Research Methods, Evidence
and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: goldsmi@mcmaster.ca

mon obstacles [1, 2]. Regardless of these difficulties, when we are presented with a RCT proclaiming that a new surgical innovation is superior to the standard approach (standard of care) we need to be confident that this is the case. Less often, we read about a new trial where the investigators found no difference between the novel and the standard approach. How can we be confident that, in the former case, the proclamation of superiority, and in the latter, the claim of “no difference” are valid?

This chapter will explain the methodological issues that pertain to sample size calculation and provide the reader with the necessary skills to appraise an RCT based on its sample size. Readers will find definitions of the bolded terms, and more information regarding the factors that influence power and sample size in [Appendix A](#).

Clinical Scenario

At the last minimal access surgery rounds, a junior surgical faculty member recommended that the service start using a preoperative ultrasound-guided Transversus Abdominis plane (TAP) block to control pain after laparoscopic surgery for colorectal cancer. She claimed that, at the center where she did her fellowship, TAP blocks were used all the time.

A cynical senior surgeon commented that such an approach is useless; it will only lengthen the procedure, enrich the anesthetist, and do nothing to improve a patient’s pain! Furthermore, he lamented that most of the conclusions presented in the surgical literature are suspect as they are based on studies with inadequate sample size.

The program director challenged the fellow to return to the rounds the following week and present evidence supporting her recommendation from a study with a large enough sample size.

A disclaimer to our readers:

The aim of this chapter was to attempt to describe the factors impacting and the calculations associated with power and sample size. The authors suggest the use of software, and the help of a trusted biostatistician collaborator when calculating sample size.

The Importance of Power and Sample Size

Sample size is incredibly important in clinical research; because of this, surgeons and surgeon-researchers should be familiar with its influencing factors. Box 29.1 summarizes the factors that inform sample size; whether one is calculating the required sample size by hand, or using statistical software, the factors in Box 29.1 will be needed.

For the community surgeon, understanding sample size is important before applying reported conclusions of studies to their practice. For academic surgeons engaged in research, this knowledge provides the foundation for a well-executed and credible clinical trial.

Box 29.1: Factors informing sample size

Factor	Further information
Outcome of interest	Page 4–5
Minimal clinically important difference (MCID)	Page 5
Standard deviation	Page 5–6
Level of significance/chance of type I error	Page 8 and Fig. 29.1
Type II error	Page 9 and Fig. 29.1
Number of tails	Pages 10
Study design	Page 10–11
Test/test statistic	Page 11–12

How the factors informing sample size presented in Box 29.1 interact can be best understood using the formula below:

$$n = f(\alpha, \beta, \delta, \sigma, D, O, Ts, Tt)$$

*n = sample size; f = function of; α = significance level or chance of Type I error; β = chance of type II error (1–β = power); δ = MCID; σ = standard deviation, D = study design, O = outcome measure, Ts = test statistic, Tt = test tails

Therefore, we can think of sample size as a function of the factors presented on the right side

of the equation. Because of this interrelationship between the factors and sample size, if we are interested in obtaining the MCID (δ), for example, we can do so by rearranging the function so that MCID (δ) is on the left side of the equation and show that δ is a different function, g , of the other terms and sample size, n . Therefore, the relationship between MCID and the other terms can be represented in a formula as:

$$\delta = g(\alpha, \beta, \sigma, D, O, Ts, Tt, n).$$

In surgical research, the clinical scenario or research question informs both the outcome(s) to be measured and the target population to be studied. In research studies, there is a primary and sometimes a secondary outcome. The primary outcome is the endpoint that is of greatest importance, while the secondary outcome is used to measure additional effects of the intervention [3]. In the above-mentioned scenario, the primary outcome would be pain following surgery with and without the use of TAP Blocks; the target population would be colorectal patients undergoing laparoscopic surgery.

Once the outcome and the target population are determined, the next step is to decide on the appropriate outcome measure (the scale or instrument) to be used. A literature review can be used to ensure that the outcome measure is: (1) relevant; (2) meaningful; (3) *reliable*; (4) *valid*; and (5) *responsive to change*. Next, one would determine the type of data measurement the outcome measure would produce (nominal, ordinal, interval, or ratio) as this has a bearing on the type of statistical analysis that will be performed. The outcome measure for the current scenario would be the Numerical Rating Scale (NRS) pain score, which is continuous (interval) in nature.

The next step is to review previous literature to determine the MCID (δ) for your specific outcome measure. The concept of MCID was introduced in 1989 by Jaeschke, Singer, and Guyatt [4, 5]. MCID is defined as the smallest difference in the outcome of interest perceived generally by patients (or physicians) as beneficial. In other words, the MCID, refers to the smallest amount of difference in an outcome measure considered meaningful

[4–7]. For surgeons, it means the smallest difference between a novel and standard approach that will persuade them to adopt the novel intervention. If no information on the MCID for your outcome measure is available, the MCID could be computed using *effect size* (ES), defined as the number of *standard deviation units* of difference or change to be expected ($ES = \delta/\sigma$). The MCID may also be referred to as the MID (Minimal Important Difference) and the terms are often used interchangeably. In this chapter, we continue to use MCID as originally described for consistency [8].

For an RCT, effect size can be calculated as the difference in the means between two chosen times, divided by the standard deviation of scores at the first time point [7]. We can also think of the effect size in terms of the degree to which a specified nonzero difference can be observed in the population under study. The larger the value of ES, the greater the degree to which manifestation of the difference is found in the population.

As noted by Cohen [9], the ES is a very important factor in determining power and required sample size. A small effect size can be expected in patients reporting minimal or no difference in the outcome of interest. Alternatively, a large effect size is expected in patients reporting large differences in the chosen outcome. Where there is nothing in the literature, we can use the suggestions by Cohen that categorizes effect sizes as proxies, i.e., small effect = 0.2; medium effect size = 0.5, or large effect size = 0.8 as proxies for our MCID [8]. The formula for calculating effect size for change over time is shown below [7]:

$$ES = \frac{[\text{Mean}(X_{\text{Initial}}) - \text{Mean}(X_{\text{Final}})]}{\sigma_{\text{baseline}}}$$

* X : *Observed Value*; X_{Initial} : *values at baseline*;
 X_{Final} : *values following treatment/procedure*;
 σ_{baseline} : *standard deviation of the baseline values*.

It is important to note that, in clinical research, it is possible to have a positive or a negative value for ES. Therefore, it is important to consider your research question when interpreting data. For example, if we are comparing Health-Related Quality of

Life (HRQOL) scores in a sample from baseline to 6-months following surgery, then we would expect to see an improvement in HRQOL, and therefore a negative ES. For example, a baseline score of 50 on the HRQOL scale SF-36 that increases to 80 at 6 months will give a negative ES based on the above formula. However, if we are measuring pain we would expect a decrease from baseline to 6 months, and therefore a positive ES; for example, a preoperative pain score of 8 on a Visual Analog Scale (VAS) of 0–10 postoperatively improves to a score of 3.

With the MCID for your specific outcome in mind, the research question is reframed into a clinical hypothesis, a statement about the expected relationship between two variables [10]. In the above scenario, the clinical hypothesis would address the suspected relationship between pain and TAP blocks (*“There will be an important decrease in pain score for those patients who had TAP Blocks, as compared to those who did not”*). The clinical hypothesis can then be stated as a statistical hypothesis, also known as the null hypothesis, which states the lack of relationship between the two variables (*There will be no difference in pain scores between groups*) [10]. The null hypothesis is essentially the hypothesis we would like to reject or nullify.

Once you have formulated your hypotheses, you must determine the best way to answer your research question. As it is often difficult, or impossible, to study everyone in a population, a subgroup (sample) of individuals from the population can be drawn [11]. The *population* is a

group of individuals within a geographic region, or institution (e.g., city, country, hospital, or school [11]). The *sample*, which is also known as the study group, comprises those individuals that will actually be a part of the study [11]. The aim of sampling is to gather a diverse group of individuals from the population so that as many individuals are represented as possible. A representative sample allows more confidence when generalizing results found within the sample back to the population of interest.

Generalizing the results from a sample back to the population is referred to as statistical inference. Generalization is subject to some errors, one of which is making a false positive statement; this is known as Type I error [5, 11] (see Fig. 29.1). Type I error occurs when the null hypothesis is rejected, when in fact, it is true. In other words, if the results of a study indicate that there is a difference between two treatment groups when in reality there is not, a Type I error was committed. The chance of a Type I error is referred to as the level of statistical significance and is denoted by the first Greek letter α [5, 11]. For example, a level of significance of 0.05 indicates that we have a 5% risk of making a conclusion that there is a difference even if there is no difference. If we feel it is very important that a Type I error is not made then we can choose a level of significance lower than 0.05, say 0.01. To reduce the chance of adopting interventions or procedures that are not effective, it is necessary for us to control our Type I error rate.

		Reality in Target Population	
		Alternative Hypothesis is True	Null Hypothesis is True
Results of Research	Reject the Null (there is a difference)	CORRECT	Type I Error (You determined there was a difference when in fact there is not)
	Did Not Reject the Null (there is no difference)	Type II Error (You determined there was no difference when in fact there is)	CORRECT

Fig. 29.1 Explanation of Type I and Type II error

The other type of error associated with making inferences from a sample to the population is known as a Type II error, represented by the second Greek letter β . It is referred to as the chance of not rejecting the null hypothesis when it is not true, or, the chance of not detecting a difference when it exists [5, 11]. Type II error is related to the concept of power, the chance of rejecting the null hypothesis when it is not true (Fig. 29.1) [5, 11]. This works out computationally to $1 - \beta$ and is the same as the chance of detecting a difference when there is a difference to detect. For example, power of 0.80 or 80% indicates that if a difference exists between two groups in a study, there is a 20% chance the study will not detect it.

Power should be measured or calculated before (a priori) a trial begins. As Farrokhyar et al. [5] explain, a priori power and sample size calculations can reduce the risk of Type II error (false-negative result). Through conducting a power calculation a priori, researchers are aware of the sample size required to reach power and can then do their best to ensure that the necessary number of participants are recruited. When power is calculated a priori, and the required sample size is met, researchers can be confident in the results of their study knowing that they had the power to detect a difference if one existed. Conversely, if power is calculated post hoc (after the completion of a study) and power was not adequate, there is little that can be done, and the researchers may have less confidence in their results. For example, after having completed a trial, researchers conclude no statistical difference in their outcome of interest between two groups; however, if the study was found to have inadequate power post hoc, these results may be false (Type II Error).

The next steps are to determine the directionality of our hypothesis, and the variability in the outcome measure of interest. Standard deviation, the dispersion of data, helps us to understand how different an individual's score is from the others [12]. Standard deviation is represented by the Greek letter σ (for the population) or s (for the sample). Information on variability of specific outcome measures, within the target population, needs to be found in the literature.

However, usually it is the variability found within a sample that is reported, not the variability of the entire population. If this information is not available, a pilot study may need to be carried out to obtain this information [13]. In addition to reviewing the literature for variability, the literature can also be used to determine if your hypothesis needs to be tested using one-tailed or two-tailed testing. A one-tail test involves uni-directional testing of the hypothesis, meaning it tests for the possibility of a relationship in one direction and ignores the possibility of a relationship in the other direction [14]. For example, if we are interested in the improvement of HRQOL with a new surgical intervention we may expect that there will only be a change in one direction (improvement). However, by doing this we may be ignoring the fact that this new intervention may be worsening HRQOL or even causing death in some situations. A two-tail test on the other hand, involves a bidirectional testing of the hypothesis and tests for the possibility of a relationship in both directions [14]. Two-tailed tests are most often used in research, as we are usually not sure of the directionality of the relationship when designing the study.

Once your null hypothesis is formulated and directionality is decided on, you must decide on the type of study design best suited to test your hypothesis. The most appropriate type of study design depends on: (1) The research question or goal of the study; (2) Whether or not an intervention or random allocation is needed; and (3) When the outcomes are to be measured and what outcome measure to use. Surgeon-researchers should be familiar with the different study designs and the hierarchy of evidence (see Chap. 5); comprehensive texts on study designs are available but beyond the scope of this chapter [15–17].

Study design can impact sample size [5]. For example, in an RCT design, where participants are assigned into different study groups using a random allocation mechanism, such as a random number generator, it will impact the overall number of participants needed. In the case of random allocation, one needs to properly determine the

overall required sample size, as well as how many participants are needed in each group so that the study has adequate power. The final required sample size is also contingent on the allocation ratio and sample loss (i.e., attrition, or loss to follow-up). Thus, the computed sample size of the study (n) needs to be increased by a factor of $1/(1-p)$ to give us a sample of size n' [18]. That is, $n' = n/(1-p)$; where p is proportion lost, and n' is rounded up to the integer multiple of the number of arms or groups in the trial [18]. More complicated study designs, such as those utilizing stratification through predetermined subgroups (e.g., grouping by recruitment site or gender), require larger sample sizes to maintain power [5]. Calculation of power and sample size for these types of study designs are beyond the scope of this chapter. If readers would like more information on subgroups, please see Chap. 30; more detailed information on the adjustments needed to calculate sample size in more advanced study designs can be found elsewhere (see Appendix C).

Generally, the analysis of the study data is dependent on the study design, the study hypothesis, the type of data collected, and the outcome measures. After we translate the study hypothesis into a statistical hypothesis, we then test for the chance that the study would have obtained a value/result that we are interested in or something more extreme using a test statistic. The choice of a statistical test, and hence the test statistic used, depends on the study hypothesis and data collected. The test statistic can be defined as a statistic computed from the data in the study and used in a statistical test to support or reject the null hypothesis. From our scenario, the hypothesis involves a comparison of the means of the NRS between two groups, the test is the independent samples t-test of the means, and the test statistic is the t-statistic. Other tests and test statistics are available to use depending on the hypothesis tested using more complicated models that consider stratification and covariate adjustment.

While there are ways of calculating power and sample size by hand, there are many software options such as commercially available PASS software from NCSS Statistical Software [19] or freely available software such as G*Power ver-

sion 3.1 [20, 21] or R [22]. In Appendix B, we have included an assortment of screenshots using G*Power to illustrate the effects that the factors summarized in Box 29.1 have on power and sample size.

In summary, the required sample size will increase if a smaller effect size is chosen, if power or standard deviation is increased, or as significance level (α), or β is decreased. However, even when calculations are done a priori, and the most valiant of efforts are made, the proper sample size cannot always be attained. In such cases, it is best for the investigators to disclose this in the limitation section of their manuscript. In most situations, if the study was performed with the appropriate methodology, it can still contribute to the literature by being combined with similar work in systematic reviews or meta-analyses (see Chap. 16: *Systematic Review and Meta-Analysis*).

Article Selection

To address the clinical dilemma described in the clinical scenario, you will need to consult the available evidence within the surgical literature. To do so, you form a clinical research question using the PICOT format (see Chap. 3: *Developing A Surgical Research Question*) and use these terms to perform a literature search (see Chap. 4: *Finding the Evidence Through Searching the Literature*). Since you have already identified your outcome, target audience, and outcome measure, completing a research question using the PICOT format is quite simple. The PICOT format terms for this scenario would be as follows:

- **Population:** Colorectal Cancer Patients undergoing laparoscopic surgery
- **Intervention:** TAP Blocks
- **Comparative Intervention:** No use of TAP Blocks
- **Outcome:** Pain
- **Time Horizon:** 24 h following surgery

Therefore, your research question would be, “In colorectal patients undergoing laparoscopic

surgery, does the use of TAP Blocks affect pain, as measured by the Numerical Rating Scale (NRS) compared to no TAP blocks at 24 h follow-ing surgery?”

You enter “Colorectal Cancer” AND “Laparoscopic surgery” AND “TAP Blocks” AND “Pain” into Cochrane Library and limit your search to the last 10 years (2013–March 2024); the search yields five published articles (see [Appendix D](#)). A title and abstract screening reveal the most appropriate article for your research question, “*Effects of preoperative ultrasound-guided transverse abdominus plane block on pain after laparoscopic surgery for colorectal cancer: A double-blind randomized controlled trial*,” by Oh et al. [23]. Of the five available studies, this study is chosen as it is the most recent study that focuses specifically on colorectal cancer patients, the comparison of two groups (patients with and without TAP Blocks), and the primary outcome is pain as measured by the NRS. The chosen article is summarized below:

Summary of the Appraised Article by Oh et al. [21]

A sample size of 25 patients per group was obtained by the authors with the assumption that there would be 30% difference in pain between patients measured using a NRS. Allowing for potential sample loss of 10%, the authors calculated they would need 28 patients per group.

The authors report no statistically significant differences between the two groups in age, body

mass index (BMI), ASA classifications, surgical duration, anesthesia duration, or comorbidities. However, there were statistically significant differences reported in sex distribution between groups (82%:17.9% M:F in the TAP group and 51.9%:48.2% M:F in the control; $p = 0.017$), and operation type (39.3%:60.7% colon:rectum in the TAP group and 66.7%:33.3% colon:rectum in the control group; $p = 0.042$). While Oh et al. [23] stated whether statistically significant differences existed, the authors did not report if differences between groups were clinically important. The inclusion of this information is common in RCTs; however, as recommended by Altman [24] it is more impactful to state if there are clinically important differences between groups. Altman [24] states that it is not good practice to infer from the lack of statistical significance that the variable under question had no effect on the outcome of interest. In other words, even though subjective assessment requires prior knowledge of the prognostic impact of the variables of interest, this should be used in place of statistical analysis to determine similarity between two randomized groups [24]. With this in mind, it would have been more effective for Oh et al. [23] to comment on the known prognostic importance of those variables compared between groups (age, BMI, ASA, surgical and anesthesia duration, and comorbidities).

There are no statistically significant differences in NRS scores at rest nor when coughing at 1 h postanesthetic recovery (PAR) although the mean score for patients in the TAP group was slightly lower than that for the control group (see [Table 29.1](#)). Postoperative results showed no sta-

Table 29.1 Pain outcomes between patient groups at four follow-up periods [23]

		Postanesthetic recovery (PAR) mean, SD	24 H postoperative (POD1) mean, SD
Pain at rest	TAP block ($n = 28$)	5.3, 1.9	4.0, 1.6
	Control ($n = 27$)	5.9, 2.0	3.9, 1.7
	p value ^a	0.24	0.87
Pain when coughing	TAP block ($n = 28$)	6.8, 1.9	5.6, 1.8
	Control ($n = 27$)	7.2, 1.7	6.1, 1.6
	p value ^a	0.38	0.32

SD standard deviation, TAP transversus abdominis plane

^aAdjusted for gender and operation

Table restructured using information from Oh et al. [23]

tistically significant differences in NRS scores at 24, 48, or 72 h after surgery. Full results are available in the original publication [19]; Table 29.1 below summarizes the data for pain at rest and when coughing for postanesthetic recovery and 24 h postoperatively. If readers wish to see full results by Oh et al. [23], including pain scores for 48 and 72 h, please see the original publication.

Please Note: The following section is an appraisal of the Oh et al. [23] article using the concepts explained thus far in this chapter.

Questions Used to Appraise

You believe that this is an appropriate article to answer the clinical dilemma posed in the scenario; however, before presenting the evidence to your supervisor and colleagues, you want to ensure its credibility. You uncover an article by Cadeddu et al. [25] that provides a set of questions to appraise a randomized controlled trial (RCT) based on power and sample size. The Cadeddu et al. [25] article has general questions on how to appraise an RCT; however, this chapter is going to focus on those questions specific to assessing power (Box 29.2).

Box 29.2: Key questions to assess power within a study [23]

1. Was a power analysis performed?
2. Was the sample size calculation detailed for the primary outcomes?
3. Is the effect size clinically relevant?
4. Would the stated difference in treatment effect result in a change in practice?
5. Is the effect size precise and consistent with your clinical experience and previously published trials?
6. If no power analysis was completed, are the results reported appropriately to estimate power?
7. Are confidence intervals included so that the estimation of the treatment effect can be determined?

Question 1: Was a Power Analysis Performed?

As power and sample size are directly related to each other, it is important to perform a power analysis a priori. After reviewing the Oh et al. [23] article, you see that no power calculation was performed. The authors state that they considered “a statistical power of 0.8”; however, they did not perform a power analysis [23]. Using the information learned above, without performing a power calculation a priori we do not know if there was an adequate sample collected to properly detect a difference if one truly occurs. This is problematic as following completion of the study it is possible that the sample size used may not be large enough to detect a difference, and therefore, the resources used for the study could have been wasted.

Question 2: Was the Sample Size Calculation Detailed for the Primary Outcome?

The primary outcome for the Oh et al. [23] study was stated as pain following surgery, on coughing postoperative day 1 (POD 1). To measure pain, the authors used the numeric rating scale (NRS) on postoperative days (PODs) up to 36 h following surgery at rest, and on coughing [23].

Oh et al. [23] do not include their sample size calculation in the published article, instead the authors reference two previous studies [26, 27]. The first study referenced by Oh et al. [23] states that pain scores, at rest, 24-h following surgery was 4.3 [26]; the next resource states that pain when coughing was two points higher than it was at rest [27]. Oh et al. [23] use this previous information to determine that the pain score when coughing would be 6.3.

The authors then state that a 30% decrease in pain scores, which, would equal 4.4, would be considered “clinically significant.” As the authors do not state where this 30% decrease was taken from you investigate and find that the Schwenk et al. [27] study mentions that their sample size was calculated to detect a 30% decrease in pain scores. The problem is that the Schwenk et al. [27] study used the Visual Analog Scale, a similar but different pain score than Oh et al. [23] and therefore you are skeptical that this decrease in

pain score would truly be clinically relevant for this sample, or pain measure.

Continuing on with their sample size calculation, Oh et al. [23] decide to set the alpha at 0.05, and set their power to 0.8, or 80%. Using these criteria, the authors determined that the study was required to have 25 individuals in each group. The authors then use a dropout rate of 10% and determined that 28 patients would be needed in each group.

Regarding the objective statement, “investigating the effects of TAP blocks on pain after surgery,” you feel that there is a lack of detail. For example, the hypothesis does not state where the expected difference would be seen, i.e., would the difference be seen between time points, within patients, or between the patient groups? Through a review of the methods section, you see that the primary objective is to compare NRS scores on coughing between groups at POD1. Table 29.2a illustrates the time points that Oh et al. [23] used for their calculation. You feel that the authors should have stated a hypothesis as “expecting a 30% difference in NRS pain scores between patient groups on coughing at POD1”, represented as $[(H-G)*100\%]/H$.

However, you also feel that a more appropriate null hypothesis would have been to hypothesize about the difference in change in pain on coughing between PAR and POD1 for the two groups, i.e., $\Delta HD - \Delta GC$ (see Table 29.2b).

However, without information on the variability of the change in pain scores from PAR to POD

1 in this population, we are unable to calculate the appropriate sample size.

Question 3: Is the Effect Size Clinically Relevant?

While reading the Oh et al. [23] article, you realize that there is no effect size given in the article. Using the information that was provided by the authors the predetermined effect size of this study would have been 0.76 or a “large” effect size. The calculation below was determined using information from the authors, for example, they assumed a mean pain score of 6.3 (standard deviation = 2.5) in coughing in the control group. They used a 30% decrease from 6.3 in the TAP group, therefore an estimated pain score of 4.4.

$$ES = \frac{[\text{Mean}(X_{\text{initial}}) - \text{Mean}(X_{\text{Final}})]}{\sigma_{\text{baseline}}}$$

$$ES = \frac{(6.3 - 4.4)}{2.5} = \frac{1.9}{2.5} = 0.76$$

Here the sample size needed to detect this effect size is 29 in each group and with a 10% drop out rate, the actual required sample size is 33 participants per group obtained using both GPower3.1 [20, 21] and PASS [19].

The actual post hoc effect size from the study on coughing at POD1 is calculated at 0.3125, closer to a small effect size:

$$ES = \frac{[\text{Mean}(X_{\text{control}}) - \text{Mean}(X_{\text{TAP Block}})]}{\sigma_{\text{control}}}$$

Table 29.2a Comparing the mean pain scores between Tap Block and Control patients across time points (analysis performed)

	NRS rest			NRS cough		
	TAP block	Control	Test	Tap block	Control	Test
PAR	A	B	A vs B	C	D	C vs D
POD1	E	F	E vs F	G	H	G vs H

Table 29.2b Comparing change in pain scores between groups at each time point (Suggested analysis)

	NRS rest			NRS cough		
	TAP block	Control	Test	Tap block	Control	Test
PAR	A	B	–	C	D	–
POD1	E	F	–	G	H	–
POD1-PAR	E-A = ΔEA	F-B = ΔFB	ΔEA vs. ΔFB	G-C = ΔGC	H-D = ΔHD	ΔGC vs ΔHD

$$ES = \frac{(6.1 - 5.6)}{1.6} = \frac{0.5}{1.6} = 0.3125$$

With the smaller observed effect size of 0.3125, a much larger sample size of 128 participants per group is needed to detect the difference in pain score of 0.5 observed on coughing at POD1. Adjusting for a 10% sample loss, the actual sample size needed is 143 participants per group.

Question 4: Would the Difference in Treatment Effect Result in a Change in Your Practice?

The authors decided to use a 30% decrease in pain as the minimally clinically important difference; however, there was no reference given for this information and a 30% decrease in pain was not found [23]. Additionally, you do not believe that their hypothesis statement was correct (as stated in Question 2). As the hypothesis statement was ambiguous, you are not confident that the results can convince you to change your practice. The decision to change practice is subjective as introducing TAP blocks into a practice would also cause a delay in the treatment of the case. Therefore, each surgeon would have to weigh the pros and cons and determine if a possible decrease in pain is worth an increase in time and resources for the case.

Question 5: Is the Effect Size Precise and Consistent with Your Clinical Experience and Previously Published Trials?

Oh et al. [23] did not state an effect size, similarly, no reference was provided for the chosen 30% decrease in pain and therefore it is difficult to determine if the variables used by the authors are justified.

Question 6: If no Power Analysis Was Completed, Are the Results Reported Appropriately to Estimate Power?

While the authors did not calculate power, the results reported by Oh et al. [23] do allow you to conduct your own power analysis:

Using the mean difference of 0.5 in pain score on coughing at POD1, standard deviation of 1.6, alpha of 0.05 (two-sided test), and 28 patients per group, the trial post hoc power calculation is 0.231.

Question 7: Are Confidence Intervals Included So That the Estimation of the Treatment Effect Can Be Determined?

Oh et al. [23] did not provide confidence intervals in their results or conclusions. Confidence intervals are used alongside p-values to describe the results seen in a study [25]. For example, while a p-value will indicate if the difference between two treatments is statistically significant, the confidence interval will indicate both the magnitude of the difference and give an idea of the interval of values where the true value is most likely to be found [25]. Confidence intervals are a very important part of research and help the reader to interpret the results of a study; for more information on confidence intervals please see Chap. 28: *Confidence Intervals*.

Resolution of the Scenario

Based on the evidence presented in the Oh et al. [23] article, and her own appraisal of the article, the Fellow informed her program director that she is retracting her original recommendation. She recommends that they repeat the study, but this time use robust methodology to find a more convincing answer. A more robust study could include some of the following changes:

- A clearly stated hypothesis that matches the research question and guides data analysis.
 - For example, to investigate the difference in change in pain scores on coughing between the control and TAP block from PAR to POD1.
- The authors should have performed a more appropriate literature search to determine criteria for clinically important differences and

attrition rates. Criteria should be used that match the sample, the procedure, and the outcome measure being used.

- That is, the 30% decrease used by Oh et al. [23] originated in a study looking at a different intervention, and a different outcome measure.
- That is, the 10% attrition rate seems to be chosen at random, there was no stated evidence to support that this attrition rate was applicable. The authors should have used previous literature or performed a feasibility study.
- The authors collect data for 24, 48 and 72 h following surgery; however, they report on the 24-h results as the primary outcome and consider the information for the other two time points as secondary outcomes.
- The authors should have utilized a biostatistician to determine a proper sample size for the outcomes that they wanted to measure.

Appendices

Appendix A: Additional Information

Qualitative Research

This chapter only outlined how to assess power and sample size in a quantitative research study. A qualitative research study is dealt with in a different manner; the authors encourage the reader to access available resources [27–31].

Definitions for Bolded Terms

- *Reliable:*
 - The degree to which a specific instrument produces consistent results, and on multiple occasions, when no evidence of change exists [30].
- *Valid:*
 - The degree that an instrument measures what it is supposed to measure [30].
- *Responsive to change:*
 - The degree to which an instrument can detect change over time [30].

Standard Deviation:

- A measure of variability providing the “average distance” of a value from the mean. The units are always expressed the same as the original value [29].
- The variation of a measure in a population can impact sample size, if the outcome being measured is similar across all members of the population than a smaller sample size is required. However, if the outcome is heterogeneous across the population you would need a larger sample size to detect differences among the participants [5]

Additional Information

- One- or two-Tail test
 - The tail refers to the end of the distribution of the test statistic.
 - A one-tail test means testing for the possibility of a relationship in only one direction and ignoring the possibility of a relationship in the other direction.

A one-tail test enables us to test the difference in a single direction with more power because of the assumption of the direction of the difference [14].

However, we also need to take into account the possibility of missing or failing to detect a difference in the other direction and hence the possibility of rejecting a promising intervention. So as tempting as it is deciding to go with a single tail test for the sole reason of having great power, it is not suggested [14].

- A two-tail test on the other hand involves a bi-directional testing of the hypothesis and tests for the possibility of a relationship in both directions. This is because in most studies, we are not sure of the directionality of the relationship and hence two-sided test are almost universal [14].
- Loss to follow-up
 - This is an important issue in every study and more so in those comparing effects of different treatments.

- Power decreases as the proportion lost to follow up (p) multiplied [1].
- Thus, the effective sample size of the study (n) needs to be increased by a factor of $1/(1-p)$ to give us a sample of size n' . That is, $n' = n/(1-p)$ [18]; where p is proportion lost, and n' is rounded up to the integer multiple of the number of arms in the trial.
- Allocation Ratio [5]
 - This term describes the ratio of participants that must be recruited into each group in a study (i.e., into the control group and into the treatment group).
 - As the ratio moves away from 1:1 (equal number of participants in each group) the sample size increases for the same power.
 - Power also decreases as the ratio moves from 1:1.

Appendix B: Using G*Power to Compute Power [30]

Figure 29.2a below shows a screen shot of G*Power with the majority of the factors summarized in Table 29.1 highlighted. G*Power defaults to the most commonly selected effect size (δ), test statistic, Type I error rate (α), and Type II error rate (β , power = $1-\beta$), but these can all be customized by the user as we have done. Additionally, the user must select the statistical test being performed and whether it is a one or two-tailed study. Figures 29.2b, 29.2c, 29.2d gives examples of a G*Power calculation and how changing various factors impacts required sample size. This chapter does not explain the three Output Parameters—Noncentrality parameter δ , Critical t and Df; for more information, readers are redirected to the G*Power manual, available for download [32]. The power probability is computed from a non-central Student's t distribution depends on DF, the critical t , and sample size to get the power probability listed as actual power.

Figure 29.2b illustrates the required sample size for a two-tailed test, using a t -test to determine significant differences between two independent groups; an ES of 0.5 (medium), an alpha of 0.05, and a Type II error rate of 0.05 or power

of 0.95 have been selected. Using this specific situation, a sample size of 210 (105 per group) would be required. In Fig. 29.2c, the ES has been changed to 0.8 with all other factors remaining the same; here you can see that the total required sample size has decreased to 84 (42 per group). The reason for the decrease in sample size is because we are indicating to G*Power that the expected difference in the outcome between the two groups is large. Therefore, it would take fewer participants to detect a large difference (0.8) than it would if the ES for that outcome between groups was smaller (0.2); an ES of 0.2 would have increased the required sample size. Last, in Fig. 29.2d, the power has been decreased to 0.80 or 80%; you can see here that keeping ES the same (0.5) and alpha the same 0.05 as in Fig. 29.2b, by decreasing power you have a smaller required sample size.

Appendix C

For more information on calculating power and sample size in complex study designs, please see:

1. Wittes J. Sample size calculations for randomized controlled trials. *Epidemiol Rev.* 2002;24:39–53.
2. Whitley E, Ball J. Statistics review 4: Sample size calculations. *Crit Care.* 2002;6:335–41.

Appendix D

Cochrane Keyword Search

“Colorectal Cancer” AND “Laparoscopic surgery” AND “TAP Blocks” AND “Pain”
Time Horizon: 2013-Mar 2024

Cochrane Search Results

1. Oh TK, Yim J, Kim J, Eom W, Lee SA, Park SC, Oh J, Park J, Park B, Kim D. Effects of preoperative ultrasound-guided transversus abdominis plane block on pain after laparoscopic surgery for colorectal cancer: A double-blind randomized controlled trial. *Surg Endosc.* 2017;31 (1):127–24.

G*Power 3.1.9.2

File Edit View Tests Calculator Help

Central and noncentral distributions Protocol of power analyses

Test family: t tests

Statistical test: Correlation: Point biserial model

Type of power analysis: A priori: Compute required sample size - given α , power, and effect size

Input Parameters

Determine => Tail(s): One

Effect size $|p|$: 0.3

α err prob: 0.05

Power ($1 - \beta$ err prob): 0.95

Output Parameters

Noncentrality parameter δ : ?

Critical t: ?

Df: ?

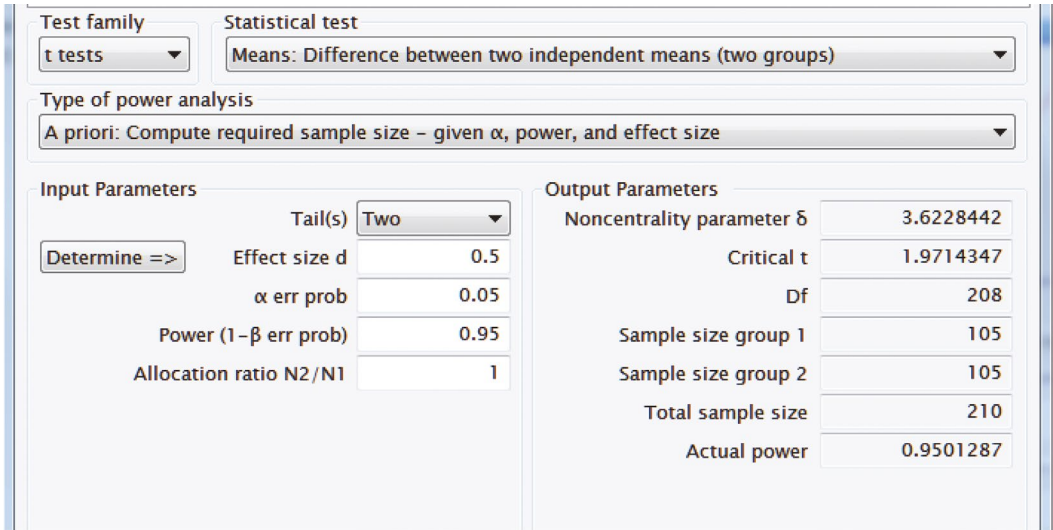
Total sample size: ?

Actual power: ?

X-Y plot for a range of values Calculate

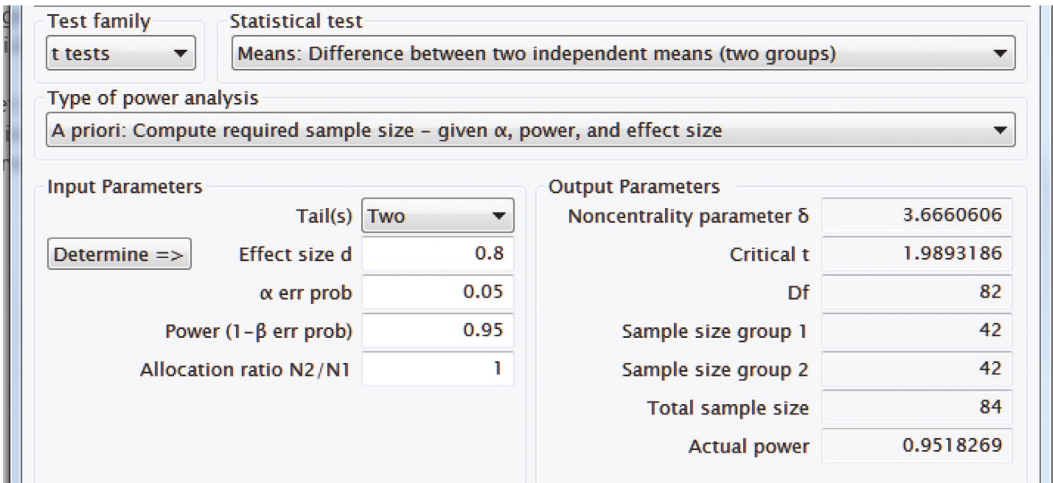
Fig. 29.2a G*Power and required factors

2. Zhao Y, Zhang HY, Yuan ZY, Han Y, Chen YR, Liu QL, Zhu T. Analgesic efficacy of postoperative bilateral, ultrasound-guided, posterior transversus abdominis plane block for laparoscopic colorectal cancer surgery: A randomized, prospective, controlled study. *BMC Anesthesiol.* 2021;21 (1):107.
3. Damadi AA, Lax EA, Smithson L, Pearlman RD. Comparison of therapeutic benefit of bupivacaine hcl transversus abdominis plane (TAP) block as part of an enhanced recovery pathway versus traditional oral and intravenous pain control after minimally invasive colorectal surgery: A prospective, random-



The screenshot shows a software interface for sample size calculation. It includes dropdown menus for 'Test family' (t tests), 'Statistical test' (Means: Difference between two independent means (two groups)), and 'Type of power analysis' (A priori: Compute required sample size - given α , power, and effect size). The 'Input Parameters' section has a 'Determine =>' button, 'Tail(s)' set to 'Two', 'Effect size d' set to 0.5, ' α err prob' set to 0.05, 'Power (1- β err prob)' set to 0.95, and 'Allocation ratio N2/N1' set to 1. The 'Output Parameters' section displays: Noncentrality parameter δ = 3.6228442, Critical t = 1.9714347, Df = 208, Sample size group 1 = 105, Sample size group 2 = 105, Total sample size = 210, and Actual power = 0.9501287.

Fig. 29.2b Example of sample size calculation [32]



This screenshot shows the same software interface as Fig. 29.2b, but with the 'Effect size d' input parameter changed to 0.8. All other input parameters remain the same. The output parameters are updated: Noncentrality parameter δ = 3.6660606, Critical t = 1.9893186, Df = 82, Sample size group 1 = 42, Sample size group 2 = 42, Total sample size = 84, and Actual power = 0.9518269.

Fig. 29.2c Example of how changing effect size impacts sample size requirement [32]

ized, double-blind trial. The American Surgeon™. 2019;85 (12):1363–1368.

4. Huang D, Song L, Li Y, Xu Z, Li X, Li C. Posteromedial quadratus lumborum block versus transversus abdominal plane block for postoperative analgesia following laparoscopic colorectal surgery: A randomized controlled trial. J Clin Anesth. 2020;62:109716.

5. Pedrazzani C, Park SY, Scotton G, Park JS, Kim HJ, Polati E, Guglielmi A, Choi GS. Analgesic efficacy of preemptive local wound infiltration plus laparoscopic-assisted transversus abdominis plane block versus wound infiltration in patients undergoing laparoscopic colorectal resection: study protocol for a randomized, multicenter, single-blind, noninferiority trial. Trials. 2019;20 (1):391.

Test family
t tests

Statistical test
Means: Difference between two independent means (two groups)

Type of power analysis
A priori: Compute required sample size – given α , power, and effect size

Input Parameters

Determine =>

Tail(s)Two

Effect size d0.5

α err prob0.05

Power ($1 - \beta$ err prob)0.8

Allocation ratio N2/N11

Output Parameters

Noncentrality parameter δ 2.8284271

Critical t1.9789706

Df126

Sample size group 164

Sample size group 264

Total sample size128

Actual power0.8014596

Fig. 29.2d Example of how changing power impacts sample size requirement [32]

References

1. Spilker B, Cramer J. Patient recruitment in clinical trials. New York: Raven Press Ltd.; 1992.

2. Thoma A, Farrokhyar F, McKnight L, Bhandari M. Practical tips for surgical research: how to optimize patient recruitment. *Can J Surg.* 2010;53(3):205–10.

3. Clinicaltrials.gov. ClinicalTrials.gov Glossary Terms [Internet]. 2024 [cited 2024 Mar 28]. Available from: <https://www.clinicaltrials.gov/study-basics/glossary>

4. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989;10(4):407–15.

5. Farrokhyar F, Reddy D, Poolman RW, Bhandari M. Why perform a priori sample size calculation? *Can J Surg.* 2013;56(3):207–13.

6. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA Guide Stat Method.* 2014;312(13):1342–3.

7. Luiz RR, Almeida RMVR. On the measurement of change in medical research. *Int J Stat Med Res.* 2012;1(2):144–7.

8. Schünemann HJ, Guyatt GH. Commentary—good-bye M(C)ID! Hello MID, where do you come from? *Health Serv Res.* 2005;40(2):593–7.

9. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale: Lawrence Erlbaum Associates; 1988.

10. LoBiondo-Wood G, Haber J. Research questions, hypotheses and clinical questions. In: *Nursing education: methods and critical appraisal for evidence-based practice.* 9th ed. Missouri: Elsevier MOSBY; 2018. p. 27–55.

11. Martinez-Mesa J, Gonzalez-Chica DA, Bastos JL, Bonamigo RR, Duqula RP. Sample size: how many

participants do I need in my research? *Bras Dermatol.* 2014;89(4):609–15.

12. University of Leicester. Measures of variability: The range, inter-quartile range and standard deviation [Internet]. 2009 [cited 2024 Mar 28]. Available from: <https://bytes.usc.edu/inf250/f23-DSci/lectures/EDA/docs/disp.pdf>

13. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, Robson R, Thabane M, Giangregorio L, Goldsmith CH. A tutorial on pilot studies: the what, why and how. *BMC Med Res Methodol.* 2010;10:1. <https://doi.org/10.1186/1471-2288-10-1>.

14. Ruxton GD, Neuhäuser M. When should we use one-tailed hypothesis testing? *Methods Ecol Evol.* 2010;1:114–7.

15. The Centre for Evidence-Based Medicine (CEBM). Study designs [Internet]. University of Oxford; 2018 [cited 2024 Mar 28]. Available from: <https://www.cebm.ox.ac.uk/resources/ebm-tools/study-designs>

16. Duke University Medical Center Library & Archives. Introduction to Evidence-Based Practice: Types of Studies [Internet]. 2024 [cited 2024 Mar 28]. Available from: <https://guides.mclibrary.duke.edu/ebm/studydesign>

17. Medical University of South Carolina Libraries. Evidence-Based Practice [Internet]. 2023 [cited 2024 Mar 28]. Available from: <https://musclibguides.com/ebp/studydesign#loaded>

18. Whitley E, Ball J. Statistics review 4: sample size calculations. *Crit Care.* 2002;6(4):335.

19. NCSS Statistical Software. Pass16. [Internet]. 2024 [cited 2024 Mar 28]. Available from <https://www.ncss.com/software/pass/>

20. Faul F, Erdfelder E, Lang A-G, Buchner A. G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods.* 2007;2007(39):175–91.

21. Faul F, Erdfelder E, Buchner A, Lang A-G. Statistical power analyses using G*power 3.1: tests for correlation and regression analyses. *Behav Res Methods*. 2009;41:1149–60.
22. The R Foundation. The R Project for Statistical Computing [Internet]. 2024 [cited 2024 Mar 28]. Available from: <https://www.r-project.org/>
23. Oh T, Yim J, Kim J, Eom W, Lee S, Park S, Oh J, Park J, Park B, Kim D. Effects of preoperative ultrasound-guided transversus abdominis plane block on pain after laparoscopic surgery for colorectal cancer: a double-blind randomized controlled trial. *Surg Endosc*. 2017;31(1):127–34.
24. Altman DG. Comparability of randomised groups. *The Statistician*. 1985;34(1):125–36.
25. Cadeddu MO, Farrokhyar F, Thoma A, Haines T, Garnett A, Goldsmith CH. Evidence-based surgery working group. Users' guide to the surgical literature: how to assess power and sample size. *Laparoscopic vs open appendectomy*. *Can J Surg*. 2008;51(6):476–82.
26. Kang SB, Park JW, Jeong SY, Nam BH, Choi HS, Kim DW, Lim SB, Lee TG, Kim D, Kim JS, Chang H, Lee HS, Kim S, Jung K, Hong Y, Kim J, Sohn D, Kim DH, Oh JH. Open versus laparoscopic surgery for mid or low rectal cancer after neoadjuvant chemoradiotherapy (COREAN trial): short-term outcomes of an open-label randomised controlled trial. *Lancet Oncol*. 2010;11:637–45.
27. Schwenk W, Bohm B, Muller JM. Postoperative pain and fatigue after laparoscopic or conventional colorectal resections. A prospective randomized trial. *Surg Endosc*. 1998;12:1131–6.
28. Malterud K, Dirk Siersma V, Dorrit Guassora A. Sample size in qualitative interview studies: guided by information power. *Innov Methods*. 2016;26(13):1753–60.
29. Sim J, Saunders B, Waterfield J, Kingstone T. Can sample size in qualitative research be determined a priori? *Int J Soc Res Methodol*. 2018;21(5):619–34.
30. Onwuegbuzie AJ, Leech NL. A call for qualitative power analyses. *Qual Quant*. 2007;41:105–21.
31. Centers for Disease Control and Prevention C. Health-Related Quality of Life (HRQOL): Measurement Properties: Validity, Reliability, and Responsiveness [Internet]. CDC; 2018 [cited 2024 Mar 28]. Available from: <https://archive.cdc.gov/#/details?url=https://www.cdc.gov/hrqol/measurement.htm>
32. Bucher A, Erdfelder E, Franz F, Lang AG. G*Power: Statistical Power Analysis for WIndows and Mac [Internet]. 2017 [cited 2024 Mar 28]. Available from <http://www.gpower.hhu.de/>



Subgroup Analyses in Surgery

30

Chloe R. Wong, Pablo E. Serrano,
Alexandra C. Hatchell,
and Sophocles H. Voineskos

Abbreviations

RCT	Randomized controlled trial
ISIS-2	Second International Study of Infarct Survival
MVC	Motor vehicle collision
ZMC	Zygomaticomaxillary complex
ORIF	Open reduction internal fixation
TXA	Tranexamic acid
POISE-3	Perioperative ischemic evaluation-3
CI	Confidence interval
ARR	Absolute risk reduction
RRR	Relative risk reduction

Introduction

As surgeons, we aim to provide the best possible care to our patients. In doing so, we collect the highest level of available evidence from the literature, such as randomized controlled trials (RCTs). Although the available evidence is usually derived from more generalized patient populations, many surgeons attempt to make treatment decisions based upon specific individual patient factors. As a result, many published RCTs perform additional analyses on specific subgroups within the defined study population.

A *subgroup* is a subset of a trial group identified on the basis of a patient or intervention char-

C. R. Wong
Department of Surgery, Division of Plastic,
Reconstructive and Aesthetic Surgery, University of
Toronto, Toronto, ON, Canada
e-mail: chloe.wong@medportal.ca

P. E. Serrano
Division of General Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence & Impact, McMaster University, Hamilton,
ON, Canada
e-mail: serrano@mcmaster.ca

A. C. Hatchell
Departments of Surgery and Oncology, Section of
Plastic Surgery, University of Calgary,
Calgary, AB, Canada
e-mail: alexandra.hatchell@medportal.ca

S. H. Voineskos (✉)
Division of Plastic, Reconstructive, and Aesthetic
Surgery, Department of Surgery, University of Toronto,
& Li Ka Shing Knowledge Institute, Unity Health,
Toronto, ON, Canada
e-mail: sophocles.voineskos@medportal.ca

acteristic that is either measured at baseline or at randomization. A *subgroup analysis* is a statistical analysis that explores whether effects of an intervention (i.e., experimental versus control) differ according to the status of a subgroup variable. Although subgroup analyses may be helpful to individualize treatment decisions and plans, these analyses may also mislead clinicians due to lack of credibility [1, 2]. Specifically, surgeons may underestimate the influence of chance on treatment effects. For example, an apparent subgroup effect of zodiac sign was reported by the investigators of the Second International Study of Infarct Survival (ISIS-2) [3]. Specifically, patients born as a Gemini or Libra who presented with myocardial infarction did not experience the same reduction in vascular mortality attributable to aspirin as patients born under other zodiac signs. Despite a statistically significant finding ($P = 0.003$ for interaction), the investigators did not believe this to be a true subgroup effect but reported these results to caution readers regarding the potential pitfalls of subgroup analyses. Therefore, when considering subgroup analyses, surgeons must be conscientious and critical in their appraisal to ensure the reported subgroup effect is meaningful and applicable.

The purpose of this chapter is to outline the criteria for rigorous subgroup analyses in methodologically sound RCTs. A clinical scenario, based upon a recent RCT in surgery, will support these criteria throughout the text.

Clinical Scenario

A 62-year-old male recently involved in a motor vehicle collision (MVC) has sustained a comminuted, depressed right zygomaticomaxillary complex (ZMC) fracture that requires open reduction internal fixation (ORIF) by your plastic surgery team. He has a past medical history significant for coronary artery disease and peripheral artery disease. You see the patient in clinic to discuss the operative plan and he asks you about bleeding. Even though you explain that bleeding is a risk of any type of surgery, and that meticulous hemostasis is obtained by means of electro-

cautery, he is nervous about the morbidity and mortality associated with bleeding. The patient asks you if you can use tranexamic acid (TXA) during the surgery as one of his friends who had hip arthroplasty recently was given TXA in his surgery to stop bleeding.

Many of your colleagues in orthopedic surgery and trauma surgery routinely use TXA to safely reduce perioperative blood loss. You therefore question if TXA use in plastic surgery reduces blood loss without an increased risk of complications (e.g., bleeding, myocardial infarction, stroke, and other thrombotic complications). You decide to perform a literature search to answer his question.

Literature Search

As stated in previous chapters, the first step to finding the best available evidence is to formulate a research question based on the PICOT format (see Chap 3: *Developing a Surgical Clinical Research Question*):

- **Population:** Patients undergoing plastic surgery
- **Intervention:** TXA
- **Comparative intervention:** Placebo
- **Outcomes:** Hemorrhage, myocardial infarction, stroke, and other thrombotic complications
- **Time:** 30 days after surgery

Using the PICOT format terms, your clinical research question could be: *“In patients undergoing plastic surgery, does TXA reduce perioperative blood loss without increasing the risk of myocardial infarction, stroke, and other thrombotic complications compared to placebo?”*

Next, you use the above terms to perform a thorough literature search (see Chap. 4: *Finding the Evidence Through Searching the Literature*): “surgical procedures, operative/AND tranexamic acid/”. To ensure you only obtain reliable and valid evidence, you limit your search with “humans and randomized controlled trial.” This search ([Appendix](#)) produces three articles, one of

which is titled “Tranexamic Acid in Patients Undergoing Noncardiac Surgery” published in the New England Journal of Medicine (NEJM) by Devereaux et al. [4]. You believe it should provide some insight into your decision for your patient with the upcoming right ZMC ORIF. You download the article for further reading and assessment.

Summary of the Appraised Article

Previous trials have not been large enough to determine if TXA increases the risk of thrombotic events in noncardiac surgery. Devereaux et al. [4] conducted an international, randomized, controlled trial (Perioperative Ischemic Evaluation-3 (POISE-3)) to answer the following questions: in patients undergoing noncardiac surgery who are at risk for bleeding and cardiovascular events, (1) does TXA reduce the incidence of life-threatening bleeding, major bleeding, or bleeding into a critical organ than placebo, and (2) is it noninferior to placebo in terms of the incidence of major cardiovascular complications within 30 days?

Patients were recruited from 114 hospitals in 22 countries. Inclusion criteria were as follows: patients must be: 45 years of age or older, undergoing inpatient noncardiac surgery, and at risk for bleeding and cardiovascular complications (known atherosclerotic disease, undergoing major surgery, an age of ≥ 70 years, and a serum creatinine level of $>175 \mu\text{mol}$ per liter). Patients were excluded if they were undergoing cardiac surgery or intracranial neurosurgery, if administration of systemic TXA during surgery was planned or if the patient had a creatinine clearance of less than 30 mL per minute or was receiving long-term dialysis.

Patients were assigned in a 1:1 ratio to receive TXA (1 g intravenous bolus) or placebo at the start and end of surgery. A total of 9535 patients were enrolled, 4757 were randomly assigned to receive TXA and 4778 were assigned to receive a placebo. The primary efficacy outcome was a composite of life-threatening bleeding, major bleeding, and bleeding into a critical organ

30 days after randomization. The primary safety outcome was a composite of myocardial injury after noncardiac surgery (e.g., myocardial infarction or isolated ischemic troponin elevation), nonhemorrhagic stroke, peripheral arterial thrombosis, and symptomatic proximal venous thromboembolism 30 days after randomization.

Within the TXA group, $n = 4729$ underwent a procedure: general surgery ($n = 1769$; 37.4%), orthopedic surgery ($n = 1083$; 22.9%), vascular surgery ($n = 699$; 14.8%), urologic surgery ($n = 598$; 12.6%), spinal surgery ($n = 237$; 5.0%), gynecologic surgery ($n = 162$; 3.4%), thoracic surgery ($n = 127$; 2.7%), low-risk surgery ($n = 39$; 0.8%), and plastic surgery ($n = 14$; 0.3%) (Table 30.1). Within the placebo group, $n = 4740$ underwent a procedure: general surgery ($n = 1773$; 37.4%), orthopedic surgery ($n = 1063$; 22.4%), vascular surgery ($n = 700$; 14.8%), urologic surgery ($n = 624$; 13.2%), spinal surgery ($n = 206$; 4.3%), gynecologic surgery ($n = 171$; 3.6%), thoracic surgery ($n = 146$; 3.1%), low-risk surgery ($n = 34$; 0.7%), and plastic surgery ($n = 23$; 0.5%). The authors also compared primary outcomes in subgroups defined by type of surgery, hemoglobin level, estimated glomerular filtration rate, and N-terminal pro-brain natriuretic peptide level.

To test the noninferiority hypothesis, the upper boundary of the one-sided 97.5% confidence interval for the hazard ratio of the composite cardiovascular outcome had to be below 1.125 and the one-sided P value had to be less than 0.025. Most outcomes were analyzed using Cox proportional-hazards models, with stratification according to blood pressure management. Hazard ratios were reported with 95% confidence intervals. A two-sided P value of less than 0.05 was considered statistically significant.

The noninferiority hypothesis was evaluated in the per-protocol population (e.g., patients who received both planned doses of TXA or placebo). For all the other analyses, data were analyzed in the intention-to-treat population (e.g., patients evaluated in the trial group to which they had been randomly assigned) (Table 30.1). A composite bleeding outcome event occurred in 433 patients (9.1%) in the TXA group and in 561

Table 30.1 Demographic characteristics and 30-day primary outcomes^a of patients undergoing noncardiac surgery (TXA vs. placebo)

	TXA (<i>n</i> = 4757)	Placebo (<i>n</i> = 4778)	Hazard ratio (95% CI)	<i>p</i> -value
Received both doses of the study drug (<i>n</i> [%])	4581 (96.3%)	4601 (96.3)		
Mean age (years $\pm$ SD)	69.5 $\pm$ 9.5	69.3 $\pm$ 9.4		
Sex (<i>n</i> [%])				
Male	2669 (56.1%)	2681 (56.1%)		
Female	2086 (43.9%)	2097 (43.9%)		
Surgery (<i>n</i> [%])				
Any procedure	4729 (99.4%)	4740 (99.2%)		
General	1769 (37.4%)	1773 (37.4%)		
Orthopedic	1083 (22.9%)	1063 (22.4%)		
Vascular	699 (14.8%)	700 (14.8%)		
Urologic	598 (12.6%)	624 (13.2%)		
Spinal	237 (5.0%)	206 (4.3%)		
Gynecologic	162 (3.4%)	171 (3.6%)		
Thoracic	127 (2.7%)	146 (3.1%)		
Low-risk	39 (0.8%)	34 (0.7%)		
Plastic	14 (0.3%)	23 (0.5%)		
Data missing on procedure performed	1 (<0.1%)	0		
No procedure performed	27 (0.6%)	35 (0.7%)		
Data missing on whether patient underwent surgery	1 (<0.1%)	3 (0.1%)		
Composite bleeding outcome ^b (<i>n</i> [%])	433 (9.1%)	561 (11.7%)	0.76 (0.67–0.87)	<0.001 ^d
Composite cardiovascular outcome ^c (<i>n</i> [%])	649 (14.2%)	639 (13.9%)	1.02 (0.92–1.14)	0.04 ^e

^aAdapted from Devereaux et al. [4]

^bThe composite bleeding outcome was a composite of life-threatening bleeding, major bleeding, and bleeding into a critical organ. Life-threatening bleeding was bleeding that was fatal, or led to: significant hypotension that requires inotrope therapy, urgent (within 24 h) surgery (other than superficial vascular repair), or intracranial hemorrhage. Major bleeding was defined as bleeding that was not specified under “life-threatening bleeding” above, and required one of the following criteria: resulted in a postoperative hemoglobin $\leq$ 70 g/L; a transfusion of $\geq$ 1 unit of red blood cells; or led to an intervention (e.g., embolization, superficial vascular repair, or nasal packing). A critical organ bleeding event was bleeding that was intracranial, intraocular, intraspinal, pericardial, retroperitoneal, or intramuscular with compartment syndrome

^cThe composite cardiovascular outcome was a composite of myocardial injury after noncardiac surgery, nonhemorrhagic stroke, peripheral arterial thrombosis, and symptomatic proximal venous thromboembolism

^dShown is the two-sided *P* value for superiority

^eShown is the one-sided *P* value for noninferiority. To show statistical significance, this *P* value had to be less than 0.025

patients (11.7%) in the placebo group (hazard ratio, 0.76; 95% confidence interval [CI], 0.67 to 0.87; absolute difference, -2.6 percentage points; 95% CI, -3.8 to -1.4 ; two-sided *P* < 0.001 for superiority) (Table 30.1). A composite cardiovascular outcome event occurred in 649 of 4581 patients (14.2%) in the TXA group and in 639 of 4601 patients (13.9%) in the placebo group (hazard ratio, 1.02; 95% CI, 0.92 to 1.14; upper

boundary of the one-sided 97.5% CI, 1.14; absolute difference, 0.3 percentage points; 95% CI, -1.1 to 1.7; one-sided *P* = 0.04 for noninferiority); the sensitivity analysis performed according to the intention-to-treat principle produced a similar result (hazard ratio, 1.03; 95% CI, 0.92 to 1.14) (Table 30.1).

These results were consistent across the pre-specified subgroups. A composite bleeding out-

Table 30.2 Thirty-day primary outcomes^a of patients undergoing noncardiac surgery (orthopedic and nonorthopedic; TXA vs. placebo)

	Orthopedic surgery				Nonorthopedic surgery			
	TXA (n = 1083)	Placebo (n = 1063)	Hazardratio (95% CI)	p-value	TXA (n = 3645)	Placebo (n = 3677)	Hazard ratio (95% CI)	p-value
Composite bleeding outcome ^b (n [%])	118 (10.9%)	156 (14.7%)	0.72 (0.57–0.92)	–	315 (8.6)	405 (11.0)	0.77 (0.67–0.90)	–
Composite cardiovascular outcome ^c (n [%])	157 (15.1)	154 (15.0)	1.00 (0.80–1.25)	–	492 (13.9)	485 (13.6)	1.03 (0.91–1.17)	–

^aAdapted from Devereaux et al. [4]

^bThe composite bleeding outcome was a composite of life-threatening bleeding, major bleeding, and bleeding into a critical organ. Life-threatening bleeding was bleeding that was fatal, or led to: significant hypotension that requires inotrope therapy, urgent (within 24 h) surgery (other than superficial vascular repair), or intracranial hemorrhage. Major bleeding was defined as bleeding that was not specified under “life- threatening bleeding” above, and required one of the following criteria: resulted in a postoperative hemoglobin ≤ 70 g/L; a transfusion of ≥ 1 unit of red blood cells; or led to an intervention (e.g., embolization, superficial vascular repair, or nasal packing). A critical organ bleeding event was bleeding that was intracranial, intraocular, intraspinal, pericardial, retroperitoneal, or intramuscular with compartment syndrome

^cThe composite cardiovascular outcome was a composite of myocardial injury after noncardiac surgery, nonhemorrhagic stroke, peripheral arterial thrombosis, and symptomatic proximal venous thromboembolism

come event occurred in 118 patients (10.9%) in the TXA group and in 156 patients (14.7%) in the placebo group within the orthopedic surgery subgroup (hazard ratio, 0.72; 95% CI 0.57 to 0.92) (Table 30.2). Similarly, a composite bleeding outcome event occurred in 315 patients (8.6%) in the TXA group and in 405 patients (11.0%) in the placebo group within the nonorthopedic surgery subgroup (hazard ratio, 0.77; 95% CI 0.67 to 0.90). In terms of composite cardiovascular outcomes, an event occurred in 157 patients (15.1%) in the TXA group and in 154 patients (15.0%) in the placebo group within the orthopedic surgery subgroup (hazard ratio, 1.0; 95% CI 0.8 to 1.25) (Table 30.2). Within the nonorthopedic surgery subgroup, a composite cardiovascular outcome event occurred in 492 patients (13.9%) in the TXA group and in 485 patients (13.6%) in the placebo group (hazard ratio, 1.03; 95% CI 0.91 to 1.17).

Appraisal of the Selected Article

You find articles by Sun and colleagues [5] and Dijkman and colleagues [6] that help you appraise the subgroup analysis in Devereaux and col-

leagues’ RCT [4]. The credibility of their subgroup analysis can now be critically analyzed on a point-by-point basis (Box 30.1).

Box 30.1: Criteria to Assess the Credibility of a subgroup Analysis

- A. Are the results valid?
1. Was the subgroup analysis based on rationale indication?
 2. Was the subgroup analysis pre-defined a priori or carried out post hoc?
 3. Was the subgroup analysis one of a small number?
 4. Did the power calculation account for between-subgroup treatment effects?
 5. Was randomization stratified for important subgroup variables?
 6. Can chance alone explain the subgroup difference? Were interaction tests used for assessing subgroup treatment effect interactions?

(continued)

Box 30.1 (continued)

7. Were the significance levels of treatment effect interactions adjusted for multiplicity?
8. Were subgroups checked for comparability of prognostic factors?
- B. What were the results?
 1. Are all performed subgroup analyses reported?
 2. Are the subgroup analyses reported as relative risk reductions?
 3. Does the emphasis of the discussion and conclusion remain on overall treatment effect?
- C. Will the results help me in caring for my patients in my practice?
 1. Is the subgroup difference consistent across other studies?
 2. Is the subgroup effect or interaction clinically important?
 3. Are the patients in the subgroup comparable to my patients?

A. Are the Results Valid?

1. *Was the subgroup based on rationale indication?*

When critically appraising a subgroup analysis, it is important to ensure that the analysis is based upon a logical rationale. If no logical rationale exists, it is difficult for a subgroup analysis to be of practical value regardless of the statistical significance of the results [1, 7]. In contrast, certain patients may be expected to have different treatment effects due to the impact of differences in risk pertaining to a specific outcome or differences in pathophysiology. In these situations, it may be appropriate to separate these patients into subgroups. In addition to the POISE-3 manuscript [4], Devereaux and colleagues also published a Rationale and Design manuscript [8]. The Rationale and Design manuscript details the purpose and methodology of the RCT, including: trial population, patient recruitment, randomization and blinding, interventions, outcomes, follow-up, data management, sample size

calculations, data analysis, trial oversight, and monitoring. This manuscript promotes transparency, reproducibility, and understanding of the trial's objectives and methodology for researchers, reviewers, and readers.

In both the POISE-3 manuscript [4] and Rationale and Design manuscript [8], Devereaux and colleagues clearly state that the subgroups were prespecified, however within the POISE-3 manuscript they do not report the rationale for the prespecified subgroups: type of surgery (orthopedic surgery vs. nonorthopedic surgery), preoperative hemoglobin level (< 120 g/L vs. ≥ 120 g/L), estimated glomerular filtration rate (eGFR; < 45 , $45- < 60$, and ≥ 60 ml min⁻¹ 1.73 m²), or N-terminal pro-brain natriuretic peptide level (NT-proBNP; < 200 , $200- < 1500$, and ≥ 1500 ng/L).

2. *Was the subgroup analysis predefined a priori or carried out post hoc?*

Subgroup analyses may be defined prior to or after the initiation of the study, but these analyses should be interpreted differently [1, 9]. Authors should identify and define specific subgroups in the protocol of the RCT a priori, including when the trial is initially registered. Furthermore, both the direction and magnitude of the hypothesized subgroups effect should also be reported a priori. These key steps avoid post hoc interpretations of the data that could bias the authors' conclusions and fit the data retroactively. In the Rationale and Design manuscript [8], Devereaux and colleagues predefined the subgroup analyses of the primary outcomes a priori as well as the direction of effect. The authors state that they expected TXA to yield greater efficacy and safety outcomes in patients undergoing orthopedic surgery, with a preoperative hemoglobin less than 120 g/L, a lower preoperative eGFR, and a higher preoperative NT-proBNP value.

Although post hoc analyses may provide important clinical information and considerations, these analyses should be approached with caution [1, 10]. Usually, post hoc analyses may be performed due to the generation of new hypotheses if the original data analyzed produces unexpected results. Since subgroup analyses are

generated by the trial data, the difference of treatment effect between subgroups may actually be due to the intervention rather than the subgroup characteristic [9, 10]. Therefore, the credibility of post hoc subgroup hypotheses is low [9, 10].

3. *Was the subgroup analysis one of a small number?*

The chance of type I error, or falsely obtaining significant subgroup effects and interactions, increases as the number of subgroup analyses increases [10–12]. The number of subgroup analyses is the product of the number of outcome analyses and the number of subgroups. To mitigate the risk of chance and sampling error, the number of subgroups should be minimized. Furthermore, subgroup analyses should be restricted to the primary outcome of the main RCT as well as the secondary outcomes applicable to specific subgroups. Ensuring subgroup analyses are both succinct and meaningful will also aid the authors in delivering a clear message to the readers regarding the findings of the study.

As proposed in the Rationale and Design manuscript [8], Devereaux and colleagues repeat the primary efficacy (composite of life-threatening, major, and critical organ bleeding at 30 days after randomization) and safety (composite of myocardial injury after noncardiac surgery, nonhemorrhagic stroke, peripheral arterial thrombosis, and symptomatic proximal venous thromboembolism at 30 days after randomization) outcomes on the subgroups of patients as per the type of surgery, hemoglobin level, estimated glomerular filtration rate, or N-terminal pro-brain natriuretic peptide level in the POISE-3 manuscript [4]. The number of outcomes analyzed is two. As such, the results have minimal risk of type I error from the limited subgroup analyses.

4. *Did the power calculation account for between-subgroup treatment effects?*

The power of a trial is defined as the probability to detect a clinically important difference between two groups, if one truly exists. Power is positively associated with the magnitude of the

treatment effect, as well as the sample size of the study. Sample size calculations for RCTs are typically based on a power of at least 80% to detect differences in outcomes between the primary treatment groups (see Chap. 29: *Power and Sample Size*). As a result, subgroup analyses are often underpowered to detect differences in treatment effect due to the much smaller number of participants within each subgroup [6, 7, 10, 13]. To detect interactions of the same size and with the same power as the overall effect, sample sizes should be inflated by four [14]. Furthermore, interaction effects are generally smaller than treatment effects, and thus even larger sample sizes would need to be obtained to demonstrate a statistically significant difference. The much larger sample sizes needed for reliable subgroup analyses are practically challenging to achieve in the clinical setting. When subgroup analyses are underpowered, the results might not be reliable and should be interpreted with caution.

In the Rationale and Design manuscript [8], Devereaux and colleagues calculate a sample size of 10,000. The large sample size is driven by the hypothesis of noninferiority of TXA on thrombotic events (safety outcome). Anticipating a placebo event rate of 10–12% for composite thrombotic events, the noninferiority margin was set at a hazard ratio of 1.125. This corresponds to an absolute increase of 1.18–1.39%. The relative 12.5% hazard increase as noninferiority margin was selected to be half that of the 25% relative hazard reduction used to establish the superiority of either aspirin or clonidine in comparison to placebo in POISE-2 [15, 16]. Although POISE-3 was designed to determine if TXA is noninferior to placebo, there exists the possibility that TXA may confer a small beneficial effect on postoperative cardiovascular events attributable to reduced bleeding. This potential benefit has been quantified to correspond to a true hazard ratio of 0.9. With these parameters, a sample size of 10,000 patients ensures $\geq 90\%$ power to test the noninferiority hypothesis. With regard to the efficacy outcome, a placebo rate of at least 7% was anticipated. A sample size of 10,000 patients is expected to yield $\geq 95\%$ power to demonstrate a relative hazard reduction of at least 30% (i.e., HR

0.70) with TXA compared to placebo in the composite of life-threatening, major, and critical organ bleeding. Further, the trial was projected to have 90% power to detect a relative hazard reduction of at least 25% (i.e., HR 0.75).

In the POISE-3 manuscript [4], Devereaux and colleagues explain that due to the Covid-19 pandemic, recruitment was stopped after 9500 patients. A sample size of 9500 was estimated to provide the RCT with a power of 90% to detect a hazard ratio of 0.80 or less for TXA compared to placebo, under the assumption that the incidence of composite bleeding events was 9% in the placebo group. This sample size provides the trial with 98% power for a noninferiority margin (hazard ratio of 1.125), assuming that the incidence of composite cardiovascular outcome events is 14.0% in the placebo group and an expected hazard ratio of 0.90, with adjustment for the partial factorial design.

The authors apply the same statistical tests for the within-subgroup analysis, necessitating the same number of patients (9500) for each subgroup to obtain the aforementioned power. Furthermore, more patients should have been included in each subgroup to account for the smaller interaction effect. These patient numbers were not achieved in the subgroups. Therefore, the probability of false-negative results was high for the subgroups and the results may not be considered as valid. As such, the conclusions drawn about differences in bleeding and cardiovascular events based on type of surgery, hemoglobin level, estimated glomerular filtration rate, or N-terminal pro-brain natriuretic peptide level should be interpreted with caution.

5. *Was randomization stratified for important subgroup variables?*

When designing an RCT with predefined subgroups, it is important to stratify randomization by the important subgroups [9]. Stratified randomization maintains the randomization of known and unknown prognostic variables between treatment groups that may influence the treatment effect [17]. Therefore, both type I and II errors may be reduced with stratification.

Stratification is especially important to consider in smaller trials, where the risk of differences in prognostic variables and type I error can be high [17]. Since subgroups are by definition smaller than the larger treatment arm groups, clinicians should not assume that prognostic variables between groups are similar at baseline unless stratification has taken place with randomization [17]. Since randomization ensures subgroups are similar with the exception of the designated treatment, valid conclusions can be made about the treatment efficacy within subgroups.

Devereaux and colleagues [4] did not perform stratified randomization according to the subgroup variable. Therefore, conclusions about the treatment efficacy within subgroup must be approached with caution.

6. *Can chance alone explain the subgroup difference? Were interaction tests used for assessing subgroup treatment effect interactions?*

Researchers and clinicians may be misled by the underappreciated potential of chance to influence the results of a trial [1, 10]. Tests for interaction can help determine whether the differences in effects may be explained by chance alone, where the null hypothesis of this test is that no difference exists in the true effect between subgroups. The lower the p-value, the more likely the null hypothesis can be rejected. Therefore, rather than just considering a threshold of significance (e.g., $\alpha < 0.05$), clinicians may opt to consider the p-value where the null hypothesis is increasingly likely to be false as the criterion level of α gets progressively smaller [12].

According to both the POISE-3 manuscript [4] and Rationale and Design manuscript [8], Devereaux and colleagues performed interaction tests to assess for statistical interactions between treatments. The Rationale and Design manuscript [8] describes that for each subgroup analysis, a Cox proportional hazards model was used to evaluate each primary outcome, incorporating a subgroup interaction term to assess subgroup effects. A subgroup effect was considered to be present if the interaction term of treatment and subgroup was statistically significant at a p-value < 0.05 .

7. *Were the significance levels of treatment effect interactions adjusted for multiplicity?*

As the number of subgroup analyses increases, there is a greater probability of detecting a positive finding caused by chance alone [9, 12]. Therefore, it is important to adjust within-subgroup treatment effects for multiplicity when multiple subgroup analyses are performed concurrently. For example, the Bonferroni method adjusts for multiplicity by dividing the overall significance level (e.g., 0.05) by the total number of subgroup analyses (e.g., 10). As a result, a new significance level for each subgroup analysis is calculated (e.g., 0.005).

In both the POISE-3 manuscript [4] and Rationale and Design manuscript [8], Devereaux and colleagues did not make adjustments for multiplicity. Therefore, readers should be aware of the potential increase in type I error risk.

8. *Were subgroups checked for comparability of prognostic factors?*

Despite randomization and stratification (if applicable), differences in prognostic variables may persist between subgroups simply due to chance [6]. As such, it is helpful to evaluate the subgroups for differences in prognostic variables after randomization, especially if those prognostic variables are believed to bias the treatment effect. If subgroups are not similar with regard to specific prognostic variables, the investigators should clearly acknowledge this to caution readers' interpretation of certain results.

In both the POISE-3 manuscript [4] and Rationale and Design manuscript [8], Devereaux and colleagues did not comment on potential differences in prognostic variables between subgroups. Therefore, this information is unknown.

B. *What Were the Results?*

1. *Are all performed subgroup analyses reported?*

Since the probability of a positive finding due to chance increases with the number of subgroup analyses performed [9, 12], the validity of a sub-

group analysis is dependent on how many other subgroup analyses were performed but not reported [9]. For instance, investigators may selectively report only statistically significant subgroup analyses [18]. As a result, clinicians may incorrectly believe in the reliability of the results and conclude that there is a difference in treatment effect (see Chap. 6: *Evaluating Surgical Interventions*).

Devereaux and colleagues registered their RCT with [ClinicalTrials.gov](https://clinicaltrials.gov) (a clinical trials registry; NCT03505723). You enter the study number into the registry website in hopes that you can review the RCT protocol. This detail is not available on the website. However, according to their Rationale and Design manuscript [8], Devereaux and colleagues conducted all planned subgroup analyses.

2. *Are the subgroup analyses reported as relative risk reductions?*

The magnitude of the treatment effect can be reported via either absolute or relative risk reduction [6]. Absolute risk reduction (ARR) refers to the difference in absolute risk of the outcome between the treatment groups. In contrast, relative risk reduction (RRR) provides an estimation of the risk that is removed by the treatment of interest.

RRR may be more appropriate to use when describing subgroup effects, as RRR tends to be more similar across risk groups compared to ARR [19, 20]. For example, the ARR of a given treatment cannot be large if the patient already has a low baseline risk. Likewise, a patient with a high baseline risk may have a large ARR following a particular treatment. Therefore, although the ARR may be different for these two groups, the RRR may be similar in value. As such, readers may conclude that there is a significant treatment effect between subgroups if ARR is only considered, when, in fact, similar RRR between subgroups indicates that no difference exists.

As per the POISE-3 manuscript [4] and Rationale and Design manuscript [8], Devereaux and colleagues reported hazard ratios for their primary and secondary outcomes, as well as for

the subgroup analyses (Tables 30.1 and 30.2). Although risk reductions were not reported for the subgroups, the reader can calculate this by using percentages from the table.

3. *Does the emphasis of the discussion and conclusion remain on overall treatment effect?*

Interestingly, industry-funded studies are more likely to perform and present subgroup analyses when the primary outcomes have nonsignificant findings [18]. Devereaux and colleagues [4] demonstrated differences in composite bleeding outcome events between the TXA and placebo groups (hazard ratio, 0.76; 95% confidence interval [CI], 0.67 to 0.87; absolute difference, -2.6 percentage points; 95% CI, -3.8 to -1.4; two-sided $P < 0.001$ for superiority), but no difference in composite cardiovascular outcome events (hazard ratio, 1.02; 95% CI, 0.92 to 1.14; upper boundary of the one-sided 97.5% CI, 1.14; absolute difference, 0.3 percentage points; 95% CI, -1.1 to 1.7; one-sided $P = 0.04$ for noninferiority).

The results of subgroup analyses are used in more than 25% of RCTs to support the conclusions of the studies [21]. However, due to the exploratory nature of subgroup analyses, discussions regarding these results should be limited and should not affect the conclusion of an RCT [2]. In their POISE-3 manuscript [4], Devereaux and colleagues remained consistent with primary objective of their study and did not use the subgroup analyses to influence the conclusion. The conclusion is clear, and the emphasis is placed on the fact that the incidence of composite bleeding was significantly lower in the TXA group compared to placebo. The difference in composite cardiovascular events between the TXA and placebo groups is small, however noninferiority of TXA was not established.

C. Will the Results Help Me in Caring for My Patients in My Practice?

1. *Is the subgroup difference consistent across other studies?*

Subgroup analyses have more credibility when the subgroup interaction effect is consistent

across different studies [10, 12]. It is however difficult to compare subgroup analyses across studies. Given that subgroups can be small in size, a lack of power and thus unreliable results can pose an issue [6]. Moreover, there is often substantial variation in surgical studies with regard to study design, populations, interventions, and outcomes. As a result, it can be challenging to compare these studies and conclude meaningful results [6].

In both the POISE-3 manuscript [4] and Rationale and Design manuscript [8], Devereaux and colleagues report their RCT to be the first large trial (114 hospitals across six continents) to evaluate the efficacy and safety of TXA in patients undergoing noncardiac surgery. Several other RCTs have demonstrated that TXA administration is effective in preventing bleeding in diverse settings, including: acute trauma [22], obstetrics [23], orthopedic surgery [24], and cardiac surgery [25]. For instance, the CRASH-2 trial investigated the effect of early administration of TXA on death and transfusion requirement in bleeding trauma patients. Investigators found a significant reduction in all-cause mortality (relative risk 0.91, 95% CI 0.85–0.97; $P = 0.0035$) and risk of death due to bleeding (relative risk 0.85, 95% CI 0.76–0.96; $P = 0.0077$).

2. *Is the subgroup effect or interaction clinically important?*

Readers may decide whether the observed subgroup effects or interactions are clinically relevant based on several factors. For instance, the subgroup analysis should be based on a rational indication, the intervention under study should be frequently provided to patients, the subgroup variables should be commonly used, and the outcomes presented should be clinically meaningful [6]. The presence of these factors contributes to the clinical meaningfulness and applicability subgroup analyses to the patient groups of the targeted audience (e.g., surgeons). Ideally, future studies should aim to replicate the findings from subgroup analyses.

In both the POISE-3 manuscript [4] and Rationale and Design manuscript [8], Devereaux

and colleagues selected the subgroups and performed the subgroup analyses with clinically important results. Hemorrhage poses a risk in all surgical procedures, and it is the responsibility of the surgeon to manage it effectively and safely. Consequently, surgeons in diverse specialties have started employing TXA in a broad spectrum of patients. For instance, patients may vary by type of surgery, hemoglobin level, estimated glomerular filtration rate, or N-terminal pro-brain natriuretic peptide level. Therefore, the authors presented subgroup analyses that were clinically important to the intended audience of surgeons.

3. *Are the patients in the subgroup comparable to my patients?*

To determine the comparability of a subgroup of patients in a study with your own group of patients, readers should carefully evaluate the patient characteristics within the subgroup of interest [6]. These characteristics are usually defined within the patient demographics, as well as the inclusion and exclusion criteria of the RCT. Due to strict inclusion and exclusion criteria of RCTs, differences are likely to exist between patients in an RCT and the patients in a reader's practice. Therefore, readers must exercise caution when drawing conclusions from a particular study and implementing them in their clinical practice.

In the clinical scenario outlined at the beginning of this article, the patient is a 62-year-old male with a right ZMC fracture who is planning to undergo ORIF. The results of Devereaux and colleagues [4] subgroup analyses of patients undergoing reconstructive surgery could be applied to the patient in the clinical scenario if he was at risk for bleeding and cardiovascular complications (e.g., known atherosclerotic disease, undergoing major surgery, an age of ≥ 70 years, and a serum creatinine level of >175 μmol per liter [2.0 mg per deciliter]) [4].

Resolution of the Scenario

After reading the article by Devereaux and colleagues [4], you conclude that the incidence of the composite bleeding outcome was significantly lower with TXA than with placebo in patients undergoing noncardiac surgery. You make this conclusion based on the overall study results, since it was a methodologically sound RCT with sufficient power to detect the overall treatment effect. While the authors state that they considered the credibility criteria to judge the reliability of a subgroup effect [16], as a reader, you remain uncertain how to evaluate the subgroup analyses due to a lack of reporting of the power calculation for between-subgroup treatment effects, stratified randomization according to the subgroup variable, adjustment for multiplicity, and assessment of differences in prognostic variables between subgroups.

Appendix

Search Strategy:

Ovid MEDLINE: Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE® Daily and Ovid MEDLINE® <1946-March 25, 2024>

1. Surgical Procedures, Operative/(56918)
2. Tranexamic Acid/(5193)
3. 1 and 2 (18)
4. Limit 3 to (humans and randomized controlled trial) (3)

Search Results:

1. Devereaux PJ, Marcucci M, Painter TW, Conen D, Lomivorotov V, Sessler DI, Chan MTV, Borges FK, Martínez-Zapata MJ, Wang CY, Xavier D, Ofori SN, Wang MK, Efremov S, Landoni G, Kleinlugtenbelt YV, Szczeklik W, Schmartz D, Garg AX, Short TG, Wittmann M, Meyhoff CS, Amir M, Torres D, Patel A,

- Duceppe E, Ruetzler K, Parlow JL, Tandon V, Fleischmann E, Polanczyk CA, Lamy A, Astrakov SV, Rao M, Wu WKK, Bhatt K, de Nadal M, Likhvantsev VV, Paniagua P, Aguado HJ, Whitlock RP, McGillion MH, Prystajecy M, Vincent J, Eikelboom J, Copland I, Balasubramanian K, Turan A, Bangdiwala SI, Stillo D, Gross PL, Cafaro T, Alfonsi P, Roshanov PS, Belley-Côté EP, Spence J, Richards T, VanHelder T, McIntyre W, Guyatt G, Yusuf S, Leslie K. Tranexamic Acid in Patients Undergoing Noncardiac Surgery. *N Engl J Med.* 2022;386(21):1986–97.
2. Wright GP, Wolf AM, Waldherr TL, Ritz-Holland D, Laney ED, Chapman HA, Lane BR, Assifi MM, Chung MH. Preoperative tranexamic acid does not reduce transfusion rates in major oncologic surgery: Results of a randomized, double-blind, and placebo-controlled trial. *J Surg Oncol.* 2020;122(6):1037–1042.
 3. Schulze S, Drenck NE, Hjortso E, Kehlet H. Influence of combined neural blockade, H1- and H2-receptor and serotonin2-receptor blockade, indomethacin and tranexamic acid on leucocyte, temperature and acute-phase protein response to surgery. *Acta Chir Scand.* 1988;154(5, 6):329–33.
-
- ## References
1. Sun X, Ioannidis JPA, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis: users' guide to the medical literature. *JAMA.* 2014;311(4):405–11.
 2. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, Bala MM, Bassler D, Mertz D, Diaz-Granados N, Vandvik PO, Malaga G, Srinathan SK, Dahm P, Johnston BC, Alonso-Coello P, Hassounah B, Walter SD, Heels-Ansdell D, Bhatnagar N, Altman DG, Guyatt GH. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ.* 2012;344:e1553.
 3. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. *Lancet.* 1988;2(8607):349–60.
 4. Devereaux PJ, Marcucci M, Painter TW, Conen D, Lomivorotov V, Sessler DI, Chan MTV, Borges FK, Martínez-Zapata MJ, Wang CY, Xavier D, Ofori SN, Wang MK, Efremov S, Landoni G, Kleinlugtenbelt YV, Szczeklik W, Schmartz D, Garg AX, Short TG, Wittmann M, Meyhoff CS, Amir M, Torres D, Patel A, Duceppe E, Ruetzler K, Parlow JL, Tandon V, Fleischmann E, Polanczyk CA, Lamy A, Astrakov SV, Rao M, Wu WKK, Bhatt K, de Nadal M, Likhvantsev VV, Paniagua P, Aguado HJ, Whitlock RP, McGillion MH, Prystajecy M, Vincent J, Eikelboom J, Copland I, Balasubramanian K, Turan A, Bangdiwala SI, Stillo D, Gross PL, Cafaro T, Alfonsi P, Roshanov PS, Belley-Côté EP, Spence J, Richards T, VanHelder T, McIntyre W, Guyatt G, Yusuf S, Leslie K. Tranexamic acid in patients undergoing noncardiac surgery. *N Engl J Med.* 2022;386(21):1986–97.
 5. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ.* 2010;340:c117.
 6. Dijkman B, Kooistra B, Bhandari M. How to work with a subgroup analysis. *Can J Surg.* 2009;52(6):515–22.
 7. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med.* 1992;116(1):78–84.
 8. Marcucci M, Painter TW, Conen D, Leslie K, Lomivorotov VV, Sessler D, Chan MTV, Borges FK, Martínez Zapata MJ, Wang CY, Xavier D, Ofori SN, Landoni G, Efremov S, Kleinlugtenbelt YV, Szczeklik W, Schmartz D, Garg AX, Short TG, Wittmann M, Meyhoff CS, Amir M, Torres D, Patel A, Duceppe E, Ruetzler K, Parlow JL, Tandon V, Wang MK, Fleischmann E, Polanczyk CA, Jayaram R, Astrakov SV, Rao M, VanHelder T, Wu WKK, Cheong CC, Ayad S, Abubakirov M, Kirov M, Bhatt K, de Nadal M, Likhvantsev V, Iglesias PP, Aguado HJ, McGillion M, Lamy A, Whitlock RP, Roshanov P, Stillo D, Copland I, Vincent J, Balasubramanian K, Bangdiwala SI, Biccari B, Kurz A, Srinathan S, Petit S, Eikelboom J, Richards T, Gross PL, Alfonsi P, Guyatt G, Belley-Cote E, Spence J, McIntyre W, Yusuf S, Devereaux PJ. Rationale and design of the PeriOperative ISchemic Evaluation-3 (POISE-3): a randomized controlled trial evaluating tranexamic acid and a strategy to minimize hypotension in non-cardiac surgery. *Trials.* 2022;23(1):101.
 9. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet.* 2005;365(9454):176–86.
 10. Guyatt GWP, Ioannidis J. When to believe a subgroup analysis. In: Guyatt GDR, Meade MO, Cook DJ, editors. *User's guides to the medical literature: a manual for evidence-based clinical practice.* 2nd ed. Toronto: McGraw-Hill; 2008.
 11. Buyse ME. Analysis of clinical trial outcomes: some comments on subgroup analyses. *Control Clin Trials.* 1989;10(4 Suppl):187s–94s.
 12. Sun X, Heels-Ansdell D, Walter SD, Guyatt G, Sprague S, Bhandari M, Sanders D, Schemitsch E, Tornetta P 3rd, Swiontkowski M. Is a subgroup claim believable? A user's guide to subgroup analysis.

- ses in the surgical literature. *J Bone Joint Surg Am*. 2011;93(3):e8.
13. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*. 1991;266(1):93–8.
 14. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol*. 2004;57(3):229–36.
 15. Devereaux PJ, Xavier D, Pogue J, Guyatt G, Sigamani A, Garutti I, Leslie K, Rao-Melacini P, Chrolavicius S, Yang H, Macdonald C, Avezum A, Lanthier L, Hu W, Yusuf S. Characteristics and short-term prognosis of perioperative myocardial infarction in patients undergoing noncardiac surgery: a cohort study. *Ann Intern Med*. 2011;154(8):523–8.
 16. Devereaux PJ, Sessler DI, Leslie K, Kurz A, Mrkobrada M, Alonso-Coello P, Villar JC, Sigamani A, Biccard BM, Meyhoff CS, Parlow JL, Guyatt G, Robinson A, Garg AX, Rodseth RN, Botto F, Buse GL, Xavier D, Chan MTV, Tiboni M, Cook DJ, Kumar PA, Forget P, Malaga G, Fleischmann E, Amir M, Eikelboom J, Mizera R, Torres D, Wang CY, VanHelder T, Paniagua P, Berwanger O, Srinathan S, Graham M, Pasin L, Manach YL, Gao P, Pogue J, Whitlock R, Lamy A, Kearon C, Chow C, Pettit S, Chrolavicius S, Yusuf S. Clonidine in patients undergoing noncardiac surgery. *N Engl J Med*. 2014;370(16):1504–13.
 17. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RJ. Stratified randomization for clinical trials. *J Clin Epidemiol*. 1999;52(1):19–26.
 18. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, Bala MM, Bassler D, Mertz D, Diaz-Granados N, Vandvik PO, Malaga G, Srinathan SK, Dahm P, Johnston BC, Alonso-Coello P, Hassounah B, Truong J, Dattani ND, Walter SD, Heels-Ansdell D, Bhatnagar N, Altman DG, Guyatt GH. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. *BMJ*. 2011;342:d1569.
 19. Furukawa TA. From effect size into number needed to treat. *Lancet*. 1999;353(9165):1680.
 20. Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med*. 1998;17(17):1923–42.
 21. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000;355(9209):1064–9.
 22. Collaborators C-t, Shakur H, Roberts I, Bautista R, Caballero J, Coats T, Dewan Y, El-Sayed H, Gogichaishvili T, Gupta S, Herrera J, Hunt B, Iribhogbe P, Izurieta M, Khamis H, Komolafe E, Marrero M-A, Mejía-Mantilla J, Miranda J, Morales C, Olaomi O, Ollidashi F, Perel P, Peto R, Ramana PV, Ravi RR, Yuthakasemsunt S. Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (CRASH-2): a randomised, placebo-controlled trial. *Lancet (London)*. 2010;376(9734):23–32.
 23. Shakur H, Roberts I, Fawole B, Chaudhri R, El-Sheikh M, Akintan A, Qureshi Z, Kidanto H, Vwalika B, Abdulkadir A, Etuk S, Noor S, Asonganyi E, Alfirevic Z, Beaumont D, Ronsmans C, Arulkumaran S, Grant A, Afsana K, Gülmezoglu M, Hunt B, Olayemi O, Roberts I, Chalmers I, Lumbiganon P, Piaggio G, Brady T, Shakur H, Roberts I, Alfirevic Z, Elbourne D, Gülmezoglu M, Ronsmans C, Balogun E, Pepple T, Prowse D, Quashi N, Barneston L, Barrow C, Beaumont D, Cook L, Frimley L, Gilbert D, Gilliam C, Jackson R, Kawahara T, Miah H, Kostrov S, Ramos M, Roberts I, Shakur H, Edwards P, Godec T, Huque S, Fawole B, Okunade O, Adetayo O, Chaudhri R, Kayani A, Javai K, Fawole B, Chaudhri R, Biryabarema C, Qureshi Z, Tchounzou R, El-Sheikh M, Kidanto H, Regmi M, Vwalika B, Dallaku K, Sahani M, Akhter S, Abdulkadir A, Meda N, Dah AK, Akintan A, Odekunle O, Monehin O, Ojo A, Akinbinu G, Offiah I, Etuk S, Akpan U, Udofia U, Okon U, Omoronyia E, James O, Olayemi O, Bello N, Adeyemi B, Aimakhu K, Akinsanya O, Adeleye B, Adeyemi O, Oluwatosin K, Aboyeji A, Adeniran A, Adewale A, Olaomo N, Omo-Aghoja L, Okpako E, Oyeye L, Alu F, Ogudu J, Ladan E, Habib I, Okusanya B, Onafowokan O, Isah D, Aye A, Okogbo F, Aigere E, Ogbiti M, Onile T, Salau O, Amode Y, Shoretire K, Owodunni A, Ologunde K, Ayinde A, Alao M, Awonuga O, Awolaja B, Adegbola O, Habeebu-Adeyemi F, Okunowo A, Idris H, Okike O, Madueke N, Mutihir J, Joseph N, Adebudo B, Fasanu A, Akintunde O, Abidoye O, Opreh O, Udonwa S, Dibia G, Bazuaye S, Ifemeje A, Umoiyoho A, Inyang-Etoh E, Yusuf S, Olayinka K, Adeyemi B, Ajenifuja O, Ibrahim U, Adamu YB, Akinola O, Adekola-Oni G, Kua P, Iheagwam R, Idrisa A, Geidam A, Jogo A, Agulebe J, Ikechebelu J, Udegbonum O, Awoleke J, Adelekan O, Sulayman H, Ameh N, Onaolapo N, Adelodun A, Golit W, Audu D, Adeniji A, Oyelade F, Dattijo L, Henry P, Adeyemi B, Loto O, Umeora O, Onwe A, Nzeribe E, Okorochukwu B, Adeniyi A, Gbejegbe E, Ikpen A, Nwosu I, Sambo A, Ladipo O, Abubakar S, Okike ON, Nduka EC, Ezenkwele EP, Onwusulu D, Irinyenikan TA, Singh S, Bariweni A, Galadanci H, Achara P, Osayande O, Gana M, Chaudhri R, Jabeen K, Mobeen A, Mufti S, Zafar M, Noor S, Ahmad B, Munawar M, Gul J, Usman N, Shaheen F, Tariq M, Sadiq N, Batool R, Ali HS, Jaffer M, Baloch A, Mukhtiar N, Ashraf T, Asmat R, Khudaiddad S, Taj G, Qazi R, Dars S, Sardar F, Ashfaq S, Majeed S, Jabeen S, Karim R, Burki F, Bukhari SR, Gul F, Jabeen M, Sherin A, Ain Q, Rao S, Shaheen U, Manzoor S, Masood S, Rizvi S, Ali A, Sajid A, Iftikhar A, Batool S, Dar L, Sohail S, Rasul S, Humayun S, Sultana R, Manzoor S, Mazhar S, Batool A, Nazir A, Tasnim N, Masood H, Khoro S, Surhio N, Aleem S, Israr N, Javed S, Bashir L, Iqbal S, Aleem F, Sohail R, Iqbal S, Dojki S, Bano A, Saba N, Hafeez M, Akram N, Israr N, Shaheen R, Hashmi H, Arshad S, Hussain

- R, Khan S, Shaheen N, Khalil S, Sachdev P, Arain G, Zarreen A, Saeed S, Hanif S, Tariq N, Jamil M, Chaudhry S, Rajani H, Wasim T, Aslam S, Mustafa N, Quddusi H, Karim S, Sultana S, Harim M, Chohan M, Salman N, Waqar F, Sadia S, Kahloon L, Manzoor S, Amin S, Akram U, Ikram A, Kausar S, Batool T, Naila B, Kyani T, Biryabarema C, Bulime R, Akello R, Lwasa BN, Ayikoru J, Namulwasira C, Komagum P, Rebecca I, Annet N, Nuulu N, Nionzima E, Bwotya R, Nankya M, Babirye S, Ngongzi J, Sanchez C, Innocent N, Anitah K, Jackson A, Ndagire E, Nanyongo C, Drametu D, Meregurwa G, Banya F, Atim R, Byaruhanga E, Felix L, Iman H, Oyiengo V, Waigi P, Wangui R, Nassir F, Soita M, Msengei R, Zubier Z, Mabeya H, Wanjala A, Mwangi H, Liyayi B, Muthoka E, Osoti A, Otara A, Ongwae V, Qureshi Z, Wanjohi V, Musila B, Wekesa K, Bosire AN, Asonganyi E, Ntem A, Njoache A, Ashu A, Simo A, Tchounzou R, Keka D, Bruno K, Ndouoya A, Saadio M, Tchana M, Gwan O, Assomo P, Mutsu V, Eric N, Foumane P, Nsem P, Fouedjio J, Fouelifack Y, Tebeu PM, Nko'ayissi G, Mbong EN, Nabag W, Desougi R, Mustafa H, Eltaib H, Umbeli T, Elfadl K, Ibrahim M, Mohammed A, Ali A, Abdelrahim S, Musa M, Awadalla K, Ahmed S, Bushra M, Babiker O, Abdullahi H, Ahmed M, Safa E, Almardi H, Rayis D, Abdelgabar SA, Alfirevic Z, Houghton G, Sharpe A, Thornton J, Grace N, Smith C, Hinshaw K, Edmundson D, Ayuk P, Bates A, Bugg G, Wilkins J, Tower C, Allibone A, Oteng-Ntim E, Kidanto H, Kazumari A, Danford A, Ngarina M, Abeid M, Mayumba K, Zacharia M, Mtove G, Madame L, Massinde A, Mwambe B, Onesmo R, Ganyaka SK, Regmi M, Gupta S, Bhatt R, Agrawal A, Pradhan P, Dhakal N, Yadav P, Karki G, Shrestha BR, Vwalika B, Lubeya M, Mumba J, Silwimba W, Hansingo I, Bopili N, Makukula Z, Kawimbe A, Lubeya MK, Mtambo W, Ng'ambi M, Dallaku K, Cenameri S, Tasha I, Kruja A, Brahimaj B, Tola A, Kaza L, Sahani M, Tshombe D, Buligho E, Paluku-Hamuli R, Kacha C, Faida K, Musau B, Kalyana H, Simisi P, Mulyumba S, Jason NK, Lubamba JR, Missumba W, Islam F, Begum N, Akhter S, Chowdhury F, Begum R, Basher F, Nargis N, Kholdun A, Jesmin S, Paul S, Segni H, Ayana G, Haleke W, Abdulkadir A, Hussien H, Geremew F, Bambara M, Somé A, Ly A, Pabakba R, Fletcher H, Samuels L, Opere-Addo H, Larsen-Reindorf R, Nyarko-Jectey K, Mola G, Wai M, El Rahman M, Basta W, Khamis H, Escobar MF, Vallecilla L, Faye GE. Effect of early tranexamic acid administration on mortality, hysterectomy, and other morbidities in women with post-partum haemorrhage (WOMAN): an international, randomised, double-blind, placebo-controlled trial. *The Lancet*. 2017;389(10084):2105–16.
24. Stowers MDJ, Aoina J, Vane A, Poutawera V, Hill AG, Munro JT. Tranexamic acid in knee surgery study—a multicentered, randomized, controlled trial. *J Arthroplast*. 2017;32(11):3379–84.
 25. Myles PS, Smith JA, Forbes A, Silbert B, Jayarajah M, Painter T, Cooper DJ, Marasco S, McNeil J, Bussi res JS, McGuinness S, Byrne K, Chan MTV, Landoni G, Wallace S. Tranexamic acid in patients undergoing coronary-artery surgery. *N Engl J Med*. 2016;376(2):136–48.

Bright Huo, Christopher J. Coroneos, Neil Smart,
Cesare Stabilini, Melissa Brouwers,
and Stavros A. Antoniou

Abbreviations

CPG	Clinical practice guidelines
GERD	Gastroesophageal reflux disease
AGREE	Appraisal of guidelines, research, and evaluation

Introduction

High quality clinical practice guidelines (CPGs) are among the highest sources of evidence according to the 6S Pyramid of Evidence [1]. CPGs are tools to support patients and clinicians in decision-making [2]. If developed rigorously, guidelines provide reliable recommendations for evidence-based practice. Furthermore, CPGs can

inform clinical policy and identify gaps for future research [2]. However, guidelines are only as valid as the process used in their development [1]. Thus, guideline appraisal is an essential skill for any surgeon as practice patterns shift with changing evidence [3]. Guideline appraisal tools have been established to support clinicians, including the Appraisal of Guidelines, Research and Evaluation (AGREE)-II tool, which has been applied internationally [4].

AGREE-II is a generic instrument for application across different disciplines in healthcare [5]. However, surgical practice is influenced by distinct characteristics from those in other healthcare domains. Unlike nonsurgical and pharmacologic interventions, surgical experience and expertise directly impact the generalizability of the benefits

B. Huo
Division of General Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: brighthuo@dal.ca

C. J. Coroneos (✉)
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
Department of Health Research Methods, Evidence,
and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: coronec@mcmaster.ca

N. Smart
Department of Gastrointestinal Surgery, Royal Devon
& Exeter Hospital, Exeter, UK

C. Stabilini
Department of Surgical Sciences (DISC),
University of Genoa, Genoa, Italy
e-mail: cesare.stabilini@unige.it

M. Brouwers
School of Epidemiology and Public Health,
Faculty of Medicine, University of Ottawa,
Ottawa, ON, Canada
e-mail: melissa.brouwers@uottawa.ca

S. A. Antoniou
Department of General Surgery, Papageorgiou
General Hospital, Thessaloniki, Greece

EAES Guidelines Committee,
Eindhoven, Netherlands

and harms associated with the surgical intervention. However, surgical quality assurance is poorly reported in conventional trials, and the definition and implementation of surgical techniques can be highly variable [6]. Additionally, surgeon-experts that develop new techniques may have invaluable contributions but also may be biased due to financial or nonfinancial competing interests [5]. These factors in addition to others highlight the gaps in existing clinical practice guideline appraisal tools for surgical guidelines [5]. Surgeons interpreting the evidence underpinning the recommendations must be able to account for not only the healthcare setting but also experience and expertise of surgeons performing the interventions in the source studies.

The AGREE-S (AGREE extension for surgery guidelines) appraisal tool was developed for the development, reporting, and critical appraisal of surgical guidelines [7]. This chapter outlines a practical approach to appraising surgical guidelines for the practicing surgeon using the AGREE-S tool, while highlighting key differences from the AGREE-II tool [4, 7].

Clinical Scenario

You are a general surgeon that performs minimally invasive surgery working in the United States. You are preparing for your new consultations and notice a 62-year-old patient that has been treated medically for gastroesophageal reflux disease (GERD) over the past 15 years. They initially trialed over-the-counter antacids before receiving a proton pump inhibitor 7 years ago to control their symptoms of acid reflux. Over the last 2 years, they have experienced worsening symptoms to the point where it is now impacting their daily quality of life. They are being referred to you to discuss surgical intervention. You recall that laparoscopic fundoplication may be indicated for this patient but had recently heard that there are multiple “wrap” techniques. You generate the following question: “What is the safest and most effective fundoplication technique for the surgical management of patients with GERD?” You decide to review the latest literature to find guideline rec-

ommendations on the most effective and safest anti-reflux procedure for this patient.

Literature Search

With limited time as a practicing surgeon, you proceed with a brief search in PubMed using the phrase “GERD surgical guideline,” which returns 274 results as of May 26, 2024. You do not have the time to sift through these results.

Recognizing this, you ask your local institution’s health sciences librarian to help you develop a focused literature search using PubMed. They advise you to start by grouping terms for your population (GERD) as well as your intervention (fundoplication), then combining them together. They also note that the “guideline” filter does not always include clinical practice guidelines comprehensively, and you decide to incorporate this into your search syntax. This produces 118 results on May 26, 2024 (Supplementary [Appendix 1](#)). You further narrow using the “in the last 5 years” filter, bringing it to 39 articles. The search process takes you 5 min altogether.

One result is titled “SAGES guidelines for the surgical treatment of gastroesophageal reflux (GERD)” [8]. In contrast to other studies, you note that it is a clinical practice guideline, and that its scope is relevant to your question. Furthermore, it is up-to-date and is relevant to you as a North American surgeon. The other articles are either not CPGs, not relevant to your clinical question, or detailed recommendations not relevant for the USA (Supplementary [Appendix 2](#)). You select the Society of American Gastrointestinal & Endoscopic Surgeons (SAGES) article to delve further and recall hearing about a surgical guideline appraisal tool named “AGREE-S” [7].

Surgical Guideline Appraisal: Agree-S

As guidelines vary in robustness and quality, clinicians reading surgical guidelines must apply at least some degree of appraisal depending on their

intent of use. AGREE-S addresses unique aspects of surgical guideline development and evidence base [7]. Each item is scored from 1 to 7, and each domain receives a scaled score based on its item scores [7]. Table 31.1 lists the AGREE-S items in contrast to AGREE-II items.

Relative to AGREE-II, unique items relevant for surgical practice have been added across domains, while less relevant items have been removed. Some domains were reformatted or exchanged with new more relevant domains for surgical guidelines (Table 31.1). For instance, referencing a protocol has been added to Domain I. Domain II focuses on comprehensive, relevant stakeholder involvement and must include a guideline methodologist. Domain IV now emphasizes the views and preferences of the target population, while resource implications and variability in surgical expertise must now be considered. Plans for guideline monitoring and updates must now be explicitly outlined in Domain 6. The following passages provide an overview of key concepts to apply when considering each item.

Scope and Purpose

Readers must identify whether the surgical guideline in question references a protocol developed a priori—i.e., prior to beginning their study. This is important, as deviations from the initial study design may introduce significant bias to the findings, and recommendations. Equally, deviations from the study protocol are sometimes warranted, and justified. In these cases, departure from an initial study plan must simply be documented and rationalized. In each guideline, fundamental methodological principles must be planned, such as key questions, evidence search and synthesis approaches, funding source(s), and stakeholder involvement [5]. Authors further mitigate bias by providing their study protocol online through guideline registries, an organizational website, or other means [5]. Guideline registration is strongly encouraged [1].

Study objectives must be stated for clarity, including the guideline's aims. These may

include improving outcomes, a diagnostic process, enhancing disease prevention, and/or informing policymakers and healthcare authorities. Equally, the health question must be planned and stated in any guideline. This typically includes explicit focus on the intended population of study, surgical intervention, and comparator. Outcomes are best defined through panel consensus, such as through the grading of their importance [9], and thus may not be included as part of the research question [5]. Readers should look for these concepts, as they impact the applicability of the content for the reader's intent.

In the SAGES guideline, you note that no protocol is explicitly identified [8]. Objectives were explicitly defined, with a focus on providing evidence-based recommendations from a surgeon and patient perspective. Key questions are listed with a population, intervention, and comparator outlined. One objective of the guideline was to review partial versus complete fundoplication for the surgical management of patients with GERD. Despite the lack of a prespecified research plan, we are well-assured that this guideline is applicable to our patient.

Stakeholders

We stress the importance for the guidelines being developed ideally by a guideline committee, with (an) expert(s) in guideline development, including a guideline methodologist. Any guideline requires a team member capable of developing a study design that follows fundamental principles of guideline methodology that can guide the panel through an appraisal of the evidence and facilitate discussion on the evidence in the context of comprehensive and diverse clinical and resource considerations. Initiatives that include these team members typically produce guidelines of higher quality [10].

Moreover, the guideline panel must include any relevant stakeholder groups for the research question at hand. Logullo and colleagues give the example of a guideline on robotic colorectal surgery for cancer—relevant stakeholders may include general and colorectal surgeons, oncolo-

Table 31.1 AGREE-S and AGREE-II instrument items [4, 7]

Domain	AGREE-S items ^a	Domain	AGREE-II items ^a
Domain 1. Scope and purpose	1. The guideline has been developed according to a protocol and the link to the protocol is provided	Domain 1. Scope and purpose	1. The overall objective(s) of the guideline is (are) specifically described
	2. The overall objective(s) of the guideline is (are) specifically described		2. The health question(s) covered by the guideline is (are) specifically described
	3. The health question(s) covered by the guideline (patient, interventions/procedures, outcomes) are specifically described		3. The population (eg: patients, public) to whom the guideline is meant to apply is
	4. The guideline was supported by a guideline development committee, including a guideline methodologist		4. The guideline development group includes individuals from all relevant professional groups
Domain 2. Stakeholders	5. The guideline development group includes individuals from all relevant professional groups and patients	Domain 2. Stakeholder involvement	5. The views and preferences of the target population (patients, public) have been sought
	6. The target users of the guideline are specifically described		6. The target users of the guideline are clearly defined
	7. Systematic methods were used to search for evidence.		7. Systematic methods were used to search for evidence
	8. The criteria for selecting the evidence are clearly described		8. The criteria for selecting the evidence are clearly described.
Domain 3. Evidence synthesis	9. The strengths and limitations of the body of evidence are clearly described	Domain 3. Rigor of development	9. The strengths and limitations of the body of evidence are clearly described
	10. The views and preferences of the target population (patients, public, etc.) have been sought.		10. The methods for formulating the recommendations are clearly described
	11. The methods for formulating the recommendations are clearly described		11. The health benefits, side effects, and risks have been considered in formulating the recommendations
	12. There is an explicit link between the recommendations and the supporting evidence		12. There is an explicit link between the recommendations and the supporting evidence
Domain 4. Development of recommendation	13. The recommendations are specific and unambiguous	Domain 4. Clarity of presentation	13. The guideline has been externally reviewed by experts prior to its publication
	14. The different options for management of the condition or health issue are clearly presented		14. A procedure for updating the guideline is provided
	15. Key recommendations are easily identifiable.		15. The recommendations are specific and unambiguous
	16. The potential resource implications of applying the recommendations have been considered		16. The different options for management of the condition or health issue are clearly
	17. The guideline considers potential variability in surgical expertise of those performing the interventions/procedures		17. Key recommendations are easily identifiable

Domain 5. Editorial Independence	18. The views of the funding body have not influenced the content of the guideline	Domain 5. Applicability	18. The guideline describes facilitators and barriers to its application
	19. Competing interests of guideline development group members have been recorded and addressed		19. The guideline provides advice or tools on how the recommendations can be put into practice
	20. The guideline describes facilitators and barriers to its application		20. The potential resource implications of applying the recommendations have been considered
	21. The guideline has been externally reviewed by clinical and methodological experts prior to its publication		21. The guideline presents monitoring or auditing criteria
Domain 6. Implementation & update	22. A procedure for updating the guideline is provided	Domain 6 editorial Independence	22. The views of the funding body have not influenced the content of the guideline
	23. The guideline provides advice and/or tools on how the recommendations can be put into practice		23. Competing interests of guideline development group members have been recorded and addressed
	24. The guideline presents monitoring and/or auditing criteria		—

^aAll items are rated on a scale of 1—Strongly disagree to 7—Strongly agree

gists, radiotherapists, pathologists, nurses, healthcare economists, surgical technology specialists with expertise in robotics, and patient representatives [5]. The inclusion of relevant non-surgeons in surgical guideline development is essential, including physicians from other specialties and additional healthcare professionals who contribute to the clinical care of these patients, experts from other disciplines including research methodologists, statisticians, health information specialists, health economists, and patient partners when applicable. For the reader, this helps to reassure that diverse perspectives have analyzed the content at hand and increases the practicality and generalizability of guideline recommendations. Depending on the intent of the CPG, decision-makers responsible for funding decisions or health/institutional service design may also be appropriate. Over or underrepresentation of key stakeholder groups may introduce bias, such as exaggerated or weak recommendations for or against surgery based on the best available evidence, patient, and resource considerations. In this context, intended users of the guideline should be planned in advance and stated explicitly. This ensures that the design of the guideline meets the needs of the users and further enables the reader to interpret the applicability of the guideline to their patient.

In the SAGES guideline, a panel was used, as well as a methodologist, though no credentials aside from guideline development experience are mentioned [8]. Their stakeholder group included general surgeons and gastroenterologists but lacked a patient representative/partner. Target users are clearly specified and include surgeons, gastroenterologists, and other physicians involved in treatment decisions regarding patients with GERD. They also mention that the guidelines are meant to provide education, inform advocacy, and describe future areas of research. Lastly, they describe that the guideline is patient-centered, and that patients can use the guideline as a source of information for discussions with their physicians. Additional stakeholders may be considered for such an initiative, with the most important group being patients given their aims. Other groups include family physicians (in the

pre- and postoperative phases, related to the management of complications). Clinical administrative leaders determining access to surgical time might also be appropriate. Aside from the omission of patient input, the stakeholder group is reasonably well-represented to answer the health question of interest, with appropriate target users in mind.

Evidence Synthesis

To answer a question, a comprehensive review of the best available evidence is needed. Whether a new or updated evidence synthesis, guidelines rely on thorough systematic reviews with transparent and deliberate methodology to ensure that the evidence used to develop recommendations is not a biased sample of the literature and is of reliable quality. This begins with a search strategy with the appropriate selection of relevant databases and a strategically devised search syntax. Generally, two or more databases should be searched to avoid missing relevant studies [1]. Close collaboration with experts in systematic searching such as health science librarians is an ideal approach, depending on their availability at a given institution. Authors must follow established methodology for screening and extracting data uncovered from this search. Readers of guidelines should look for the involvement of an evidence synthesis lead with experience in systematic review and guideline development. Reviewing the search syntax is equally necessary, which is often included in a supplementary appendix or in the initial study protocol. Though the SAGES guidelines lack protocol, they do include the search syntax in their separate systematic review publication, which queried multiple databases [11].

Acknowledging the limitations of the best available evidence is equally important in the absence of high-quality evidence, and readers must be aware of the potential for absent or limited data to influence the findings, methods, or even scope of a given guideline. The use of the Grading of Recommendations, Assessment, Development, and Evaluations (GRADE)

approach to the certainty of evidence assessment is recommended [12].

In the SAGES guideline, the use of additional databases such as SCOPUS may have identified articles published in non-English journals and would be the ideal scenario when developing conventional clinical practice guidelines [8]. Authors cite robust methodology in evidence synthesis with respect to screening and data extraction. Overall, readers are practically reassured by the thoroughness of the methods applied and evidence gathered.

Development of Recommendations and Editorial Independence

Once compiled, authors must describe their methodology in moving from the best available evidence to recommendations. Details surrounding the approach to deciding on the strength and direction of recommendations are paramount [12]. Furthermore, data must be interpreted in context. As such, the involvement of patient and public views is integral as guideline authors move from evidence to recommendations. For recommendations impacting patient care, it is recommended that patients participate from the beginning, as part of the guideline development group. Their contribution as panel members aids to provide unique and diverse perspectives and plays a key role in the outcome definition and in the development of recommendations to ameliorate the consideration of sociocultural acceptability, equity, and resource considerations [5]. Furthermore, patient perspective is essential in the interpretation of health benefits, side effects, and risks of a given intervention, with consideration of the alternative options available. A systematic approach to developing recommendations is required, such as the evidence-to-decision framework using GRADE methodology [12]. Whichever approach is used, readers should expect authors to detail their methodology with respect to the capturing of these perspectives and any approach to gathering public feedback once recommendations are developed. Recommendations must be clear and unambiguous to readers, and if focused on the management of a

condition, any pertinent options should be contrasted. Moreover, it is paramount that authors maintain autonomy in the development of guideline recommendations in the presence of funding bodies. Similarly, any potential bias from conflict(s) of interest among authors must be declared from the beginning of guideline development, and any approach taken to address conflicts, if present, must be outlined.

Authors reported the use of GRADEPro Evidence to Decision tables to capture the balance of the magnitude and direction of desirable and undesirable effects of the interventions, the certainty of the evidence, and variation in values related to outcomes [8]. For the last point, they report that panel members “used their collective patient experience to approximate judgments about patient values and preferences as well associated variability.” It is established that clinicians do not accurately approximate the values and preferences of patients [13, 14], further emphasizing the importance of including patient input throughout the guideline development process. Moreover, authors state that given the patient-surgeon perspective that was applied, and not a population perspective, recommendations were developed without considering the resources required, cost-effectiveness, and equity. Acceptability and feasibility are not mentioned, but these collectively comprise essential components of the GRADE methodology and evidence-to-decision framework [12, 15]. The resulting recommendations are indeed explicitly outlined, with various options for management. Variation in surgical expertise of those performing the interventions/procedures was not considered. Though authors transparently identify key aspects of GRADE methodology that were not considered, we are assured that clinicians on the guideline panel estimated patient values based on their prior experiences to develop recommendations based on the balance of benefits and harms, as well as the certainty of the evidence. They recommend a shared decision-making approach, as patients valuing reflux over dysphagia may elect for complete fundoplication while those aiming to mitigate postoperative dysphagia would elect for partial fundoplication.

Implementation and Update

From the planning stage, guideline authors must strategize their plans for dissemination once recommendations are generated. This includes the barriers and facilitators to their implementation in surgical practice. It is important that readers are given guidance on how to implement guideline recommendations, which may include but are not limited to guideline summaries, how-to manuals, decision aids, and smartphone applications [5]. Moreover, an effective plan for monitoring, auditing, and updating the guideline must be in place to ensure that recommendations stay relevant for dynamic practice environments in surgery.

The SAGES guidelines state that implementation will be feasible, though potential barriers are not mentioned [8]. One such barrier may include a lack of expertise among surgeons for performing partial fundoplication, which may limit patient access to this intervention. Regarding updates and monitoring, authors report that repeat literature searches will be performed on a “frequent interval” though this time period is unspecified. They further state that a formal update will be undertaken when substantive new literature is identified. We are well assured that conflicts of interest were collected at the beginning of the guideline, and that funding bodies did not influence the development of the guideline.

Clinical Scenario Resolution

After finding and appraising the surgical guideline, you are prepared to have an evidence-based, patient-centered discussion about the risks and benefits of Nissen versus Toupet fundoplication. You do have expertise in both fundoplication techniques and can offer either to your patient.

You estimate that your operative time for both procedures is reasonably comparable. After discussing the risks and benefits of both approaches, your patient informs you that they prefer to mitigate the risk of dysphagia. Together, you elect for a Toupet repair. You deliver a Divisional Grand Rounds 2 months later about the evidence behind the surgical management for GERD to improve the implementation of evidence-based surgical practice, tailored to the workflow and culture of patients at your institution.

Conclusion

Surgical guideline interpretation and appraisal is an exciting and promising area of research. The advent of a new appraisal instrument named AGREE-S can support surgeons in considering unique, crucial aspects of guideline development that impact both the internal and external validity of surgical guideline recommendations. Surgeons must always remember to use clinical practice guidelines as a tool to support surgical decision-making as we move toward evidence-based, patient-centered surgical care.

Appendices

Appendix 1: Literature Search Syntax

```
(((((gastroesophageal reflux[tw]) OR (gastro-
esophageal reflux[MeSH Major Topic]))
OR (gerd[tw])) OR (gerd[MeSH Major
Topic])) AND (((((fundoplication[MeSH
Major Topic]) OR (fundoplication[tw]))
OR (Nissen[tw])) OR (Toupet[tw])) OR
(Dor[tw]))) AND ((guidelines[tw]) OR
(guideline[tw]))
```

Appendix 2: Articles Identified by the Literature Search

Authors	Year	Title
Yadlapati R, Gyawali CP, Pandolfino JE; CGIT GERD Consensus Conference Participants	2022	AGA clinical practice update on the personalized approach to the evaluation and management of GERD: Expert review
Rodríguez de Santiago E, Albéniz E, Estremera-Arevalo F, Teruel Sanchez-Vegazo C, Lorenzo-Zúñiga V	2021	Endoscopic anti-reflux therapy for gastroesophageal reflux disease
Slater BJ, Collings A, Dirks R, Gould JC, Qureshi AP, Juza R, Rodríguez-Luna MR, Wunker C, Kohn GP, Kothari S, Carlsion E, Worrell S, Abou-Setta AM, Ansari MT, Athanasiadis DI, Daly S, Dimou F, Haskins IN, Hong J, Krishnan K, Lidor A, Little V, Iow D, Petrick A, Soriano IS, Thosani N, Tyberg A, Velanovich V, Vilallonga R, Marks JM	2023	Multi-society consensus conference and guideline on the treatment of gastroesophageal reflux disease (GERD)
Slater BJ, Dirks RC, McKinley SK, Ansari MT, Kohn GP, Thosani N, Qumseya B, Billmeier S, Daly S, Crawford C, P Ehlers A, Hollands C, Palazzo F, Rodriguez N, Train A, Wassenaar E, Walsh D, Pryor AD, Stefanidis D	2021	SAGES guidelines for the surgical treatment of gastroesophageal reflux (GERD)
Xiao YL	2022	Gastroesophageal reflux disease: When East meets West
Fuchs KH, Breithaupt W, Varga G, Babic B, Schulz T, Meining A	2022	Primary laparoscopic fundoplication in selected patients with gastroesophageal reflux disease
Mohr C, Ciomperlik H, Dhanani N, Olavarria OA, Hannon C, Hope W, Roth S, Liang MK, Holihan JL	2022	Review of SAGES GERD guidelines and recommendations
Launois C, Mulette P, Ancel J, Dury S, Hagenburg J, Lebagry F, Dormoy V, Deslee G, Perotin JM	2021	[Treatment of GERD in asthma]
Drug VL, Antoniu S, Oana BB, Arghir OC, Bancila I, Bataga S, Brisc C, Cijevschi-Prelipcean C, Ciocirlan M, Ciortescu I, David L, Deleanu OC, Diculescu M, Dimitriu A, Dobru D, Dumitru E, Gheonea DI, Gheorghe C, Goldis A, Jinga M, Man M, Mateescu B, Manuc M, Mihai C, Mihaltan F, Mihaescu T, Nedelcu L, Negreanu L, Pop CM, Rajnoveanu R, Saftoiu A, Seicean A, Sporea I, Stanciu C, Surdea-Blaga T, Tantau M, Todea D, Trifan AV, Ulmeanu R, Iov DE, Dumitrascu DL	2022	Romanian guidelines for the diagnosis and treatment of GERD-induced respiratory manifestations
Morice D, Elhassan HA, Myint-Wilks L, Barnett RE, Rasheed A, Collins H, Owen A, Hughes K, Mcleod R	2022	Laryngopharyngeal reflux: Is laparoscopic fundoplication an effective treatment?
Simadibrata DM, Ngadiono E, Sinuraya FAG, Damara I, Fass R, Simadibrata M	2023	Diagnostic accuracy of gastroesophageal reflux disease questionnaire for gastroesophageal reflux disease: A systematic review and meta-analysis
Hajjar A, Verhoeff K, Jogiat U, Mocanu V, Birch DW, Switzer NJ, Wong C, Karmali S	2023	Endoscopic plication compared to laparoscopic fundoplication in the treatment of gastroesophageal reflux disease: a systematic review and meta-analysis
Ota K, Takeuchi T, Higuchi K, Sasaki S, Mori Y, Tanaka H, Hakoda A, Sugawara N, Iwatsubo T, Nishikawa H.	2024	Frontiers in endoscopic treatment for gastroesophageal reflux disease
Grupo de Trabajo de Reflujo Gastroesofágico del Comité Nacional de Gastroenterología Pediátrica	2021	[Recommendations for the diagnosis and treatment of gastroesophageal reflux in childhood]

(continued)

Authors	Year	Title
Ng J, Friedmacher F, Pao C, Charlesworth P	2021	Gastroesophageal reflux disease and need for antireflux surgery in children with cystic fibrosis: A systematic review on incidence, surgical complications, and postoperative outcomes
van Lennep M, Gottrand F, Faure C, Omari TI, Benninga MA, van Wijk MP, Krishnan U	2022	Management of gastroesophageal reflux disease in esophageal atresia patients: A cross-sectional survey amongst international clinicians
Castagneto-Gissey L, Russo MF, D'Andrea V, Genco A, Casella G	2023	Efficacy of sleeve gastrectomy with concomitant hiatal hernia repair versus sleeve-fundoplication on gastroesophageal reflux disease resolution: Systematic review and meta-analysis
Shiu SI, Chang CH, Tu YK, Ko CW	2022	The comparisons of different therapeutic modalities for idiopathic achalasia: A systematic review and network meta-analysis
Kermansaravi M, Parmar C, Chiappetta S, Shikora S, Aminian A, Abbas SI, Angrisani L, Bashir A, Behrens E, Bhandari M, Clapp B, Cohen R, Dargent J, Dilemans B, De Luca M, Haddad A, Gawdat K, Elfawal MH, Himpens J, Huang CK, Husain F, Kasama K, Kassir R, Khan A, Kow L, Kroh M, Lakdawala M, Corvala JAL, Miller K, Musella M, Nimeri A, Noel P, Palermo M, Poggi L, Poghosyan T, Prager G, Prasad A, Alqahtani A, Rheinwalt K, Ribeiro R, Shabbir A, Torres A, Villalonga R, Wang C, Mahawar K, Zundel N	2023	Best practice approach for redo-surgeries after sleeve gastrectomy, an expert's modified Delphi consensus
Hojo M, Nagahara A, Hahm KB, Iwakiri R, Watanabe T, Rani AA, Zhu Q, Chan FKL, Sollano JD, Kamiya T, Yamaguchi S, Motoya S, Fock KM, Fukudo S, Kachintorn U, Suzuki H, Murakami K; The International Gastroenterology Consensus Symposium Study Group	2020	Management of Gastroesophageal Reflux Disease in Asian countries: Results of a Questionnaire Survey
Akcelik A, Miller C, Bakhos C, Abbas A, Petrov R	2023	Endoscopic interventions in the management of the gastroesophageal reflux: a narrative review
Latha Kumar A, Sadagopan A, Mahmoud A, Begg M, Tarhuni M, N Fotso M, Gonzalez NA, Sanivarapu RR, Osman U, Nath TS	2023	Comparison of the clinical efficacy, safety, and postoperative outcomes between peroral esophageal myotomy and laparoscopic Heller's myotomy with fundoplication: A systematic review
Markar S, Andreou A, Bonavina L, Florez ID, Huo B, Kontouli KM, low DE, Mavridis D, Maynard N, Moss A, Pera M, Savarino E, Siersema P, Sifrim D, Watson DI, Zaninotto G, Antoniou SA	2022	UEG and EAES rapid guideline: Update systematic review, network meta-analysis, CINEMA and GRADE assessment, and evidence-informed European recommendations on surgical management of GERD
Mohr C, Ciomperlik H, Dhanani N, Olavarria OA, Hannon C, Hope W, Roth S, Liang MK, Holihan JL.	2023	Evaluation of ICARUS guidelines and recommendations not supported by randomized controlled trials
Schizas D, Mastoraki A, Papoutsis E, Giannakoulis VG, Kanavidis P, Tsilimigras D, Ntourakis D, Lyros O, Liakakos T, Moris D	2020	LINX(®) reflux management system to bridge the "treatment gap" in gastroesophageal reflux disease: A systematic review of 35 studies
Chiappetta S, de Falco N, Lainas P, Kassir R, Valizadeh R, Kermansaravi M	2023	Safety and efficacy of Roux-en-Y gastric bypass as revisional bariatric surgery after failed anti-reflux surgery: a systematic review
Aiolfi A, Nosotti M, Matsushima K, Perali C, Ogliari C, Del Papa N, Bonitta G, Bona D	2021	Surgical treatment of recalcitrant gastroesophageal reflux disease in patients with systemic sclerosis: a systematic review

(continued)

Authors	Year	Title
Khor YH, Bissell B, Ghazipura M, Herman D, Hon SM, Hossain T, Kheir F, Knight SL, Kreuter M, Macrea M, J Mammen M, Molina-Molina M, Selman M, Wijsenbeek M, Raghu G, Wilson KC	2022	Antacid medication and antireflux surgery in patients with idiopathic pulmonary fibrosis: A systematic review and meta-analysis
Sobral J, Machado M, Barbosa JP, Barbosa J	2024	Achalasia: Laparoscopic Heller myotomy with fundoplication versus peroral endoscopic myotomy—a systematic review and meta-analysis
Liu M, Stefanova DI, Finnerty BM, Schnoll-Sussman FH, Katz PO, Fahey TJ 3rd, Zarnegar R	2022	The impact of pneumoperitoneum on esophagogastric junction distensibility during anti-reflux surgery
Hong L, Xu K, Zhu D, Sun L, Dipasquale V, Romano C	2021	Nutritional management of children with neurological impairment in China: Current status and future directions
Depypere L, Van Veer H, Nafteux P, Coosemans W	2023	Current practice in antireflux and hiatal hernia surgery: Exploration of the Belgian field
Valinoti AC, Angeramo CA, Dreifuss N, Herbelli FAM, Schlottmann F	2024	Magnetic sphincter augmentation device for gastroesophageal reflux disease: effective, but postoperative dysphagia and risk of erosion should not be underestimated. A systematic review and meta-analysis
Lechien JR, Dapri G, Dequanter D, Rodriguez Ruiz A, Marechal MT, De Marrez LG, Saussez S, Fisichella PM	2019	Surgical treatment for laryngopharyngeal reflux disease: A systematic review
ARROW Study Group; Walker R, Wiggins T, Blencowe NS, Findlay JM, Wilson M, Currie AC, Hornby S, Markar SR, Rahman S, Lloyd M, Hollyman M, Jaunoo S	2021	A multicenter prospective audit to investigate the current management of patients undergoing anti-reflux surgery in the UK: Audit & review of anti-reflux operations & workup
Weusten BLAM, Barret M, Bredenoord AJ, Familiari P, Gonzalez JM, van Hooft JE, Lorenzo-Zúñiga V, Louis H, Martinek J, van Meer S, Neumann H, Pohl D, Prat F, von Renteln D, Savarino E, Sweis R, Tack J, Tutuian R, Ishaq S	2020	Endoscopic management of gastrointestinal motility disorders—part 2: European Society of Gastrointestinal Endoscopy (ESGE) Guideline
Martines G, Musa N, Aquilino F, Picciariello A, Altomare DF	2020	Sleeve gastrectomy combined with Nissen fundoplication as a single surgical procedure, is it really safe? A case report
Davidson JR, Franklin D, Kumar S, Mohammadi B, Dawas K, Eaton S, Curry J, De Coppi P, Abbassi-Ghadi N	2020	Fundoplication to preserve allograft function after lung transplant: Systematic review and meta-analysis
Mattioli G, Cipriani MS, Barone G, Palo F, Arrigo S, Gandullia P, Avanzini S, Wong MCY	2024	Pediatric nutritional surgery and its implications: Results from a unicentric retrospective analysis

References

- Santangelo T, Novosel LC, Cook BG, Gapsis M. Using the 6S pyramid to identify research-based instructional practices for students with learning disabilities. *Learn Disabil Res Pract.* 2015;30:91–101. <https://doi.org/10.1111/ldrp.12055>.
- Huo B, Smart N, Antoniou SA. Proposing a new surgical evidence ecosystem. *Color Dis.* 2023;25:1947–8.
- Coroneos CJ, Voineskos SH, Cornacchi SD, Goldsmith CH, Ignacy TA, Thoma A. Users' guide to the surgical literature: how to evaluate clinical practice guidelines. *Can J Surg.* 2014;57:280–6. <https://doi.org/10.1503/cjs.029612>.
- Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, Fervers B, Graham ID, Grimshaw J, Hanna SE, Littlejohns P, Makarski J, Zitzelsberger L. AGREE II: advancing guideline development, reporting and evaluation in health care. *CMAJ.* 2010;182:E839. <https://doi.org/10.1503/cmaj.090449>.
- Logullo P, Florez ID, Antoniou GA, Markar S, López-Cano M, Silecchia G, Tsokani S, Mavridis D, Brouwers M, Antoniou SA, Dayem SA, Amer Y, Bertolaccini L, Alonso-Coello P, Akl EA, Chand M, Como JJ, de Borst GJ, Di Saverio S, Emile S, Eom BW, Gorter R, Hanna G, Immonen K, Lai Q, Lumen N, Mathew JL, Montendori A, Moya M, Pellino G, Sanabria A, Saratzis A, Smart N, Stefanidis D,

- Zaninotto G. AGREE-S: AGREE II extension for surgical interventions—united European gastroenterology and European Association for Endoscopic Surgery methodological guide. *United European Gastroenterol J.* 2022;10:425–34. <https://doi.org/10.1002/ueg2.12231>.
6. Huo B, Andreou A, Onos L, Francis NK, Antoniou SA. Methods of quality assurance in multicenter trials in laparoscopic fundoplication for gastroesophageal reflux disease. *Surg Endosc.* 2023;37:6711–7.
 7. Antoniou SA, Florez ID, Markar S, Logullo P, López-Cano M, Silecchia G, Antoniou GA, Tsokani S, Mavridis D, Brouwers M, Dayem YSA, Bertolaccini L, Alonso-Coello P, Akl E, Chand M, Como JJ, de Borst GJ, Di Saverio S, Emile S, Eom BW, Gorter R, Hanna G, Immonen K, Lai Q, Lumen N, Mathew JL, Montendori A, Moya M, Pellino G, Sanabria A, Saratzis A, Smart N, Stefanidis D, Zaninotto G. AGREE-S: AGREE II extension for surgical interventions: appraisal instrument. *Surg Endosc.* 2022;36:5547–58. <https://doi.org/10.1007/s00464-022-09354-z>.
 8. Slater BJ, Dirks RC, McKinley SK, Ansari MT, Kohn GP, Thosani N, Qumseya B, Billmeier S, Daly S, Crawford CP, Ehlers A, Hollands C, Palazzo F, Rodriguez N, Train A, Wassenaar E, Walsh D, Pryor AD, Stefanidis D. SAGES guidelines for the surgical treatment of gastroesophageal reflux (GERD). *Surg Endosc.* 2021;35:4903–17. <https://doi.org/10.1007/s00464-021-08625-5>.
 9. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, Alderson P, Glasziou P, Falck-Ytter Y, Schünemann HJ. GRADE guidelines: 2. Framing the question and deciding on important outcomes. *J Clin Epidemiol.* 2011;64:395–400. <https://doi.org/10.1016/j.jclinepi.2010.09.012>.
 10. Antoniou SA, Tsokani S, Mavridis D, López-Cano M, Antoniou GA, Stefanidis D, Francis NK, Smart N, Muysoms FE, Morales-Conde S, Bonjer HJ, Brouwers MC. Guideline assessment project: filling the GAP in surgical guidelines: quality improvement initiative by an international working group. *Ann Surg.* 2019;269:642–51.
 11. McKinley SK, Dirks RC, Walsh D, Hollands C, Arthur LE, Rodriguez N, Jhang J, Abou-Setta A, Pryor A, Stefanidis D, Slater BJ. Surgical treatment of GERD: systematic review and meta-analysis. *Surg Endosc.* 2021;35:4095–123.
 12. GRADE guidelines: 3. Rating the quality of evidence. <https://www.sciencedirect.com/science/article/abs/pii/S089543561000332X>.
 13. Downey L, Au DH, Curtis JR, Engelberg RA. Life-sustaining treatment preferences: matches and mismatches between patients' preferences and clinicians' perceptions. *J Pain Symptom Manag.* 2013;46:9–19. <https://doi.org/10.1016/j.jpainsymman.2012.07.002>.
 14. Button E, Cardona M, Huntley K, Gavin NC, Leblanc TW, Olsen A, Smith M, Yates P. Clinicians' understanding of preferences and values of people with hematological malignancies at the end of life: concurrent surveys. *J Palliat Med.* 2022;25:1386–97. <https://doi.org/10.1089/jpm.2021.0490>.
 15. Granholm A, Alhazzani W, Møller MH. Use of the GRADE approach in systematic reviews and guidelines. *Br J Anaesth.* 2019;123:554–9.



Network Meta-analyses in Surgical Trials

32

Lily J. Park, Kevin Um, Dena Zera,
Emmanuelle Duceppe, and Pablo E. Serrano

Abbreviations

NMA	Network meta-analysis
GRADE	Grading of Recommendations Assessment, Development, and Evaluation
SSI	Surgical site infections
RR	Relative risk
RCT	Randomized controlled trial
SUCRA	Surface under the cumulative ranking curve
IF	Inconsistency Factor
SIDE	Separating Indirect from Direct Evidence
CI	Confidence Intervals

Introduction

Network meta-analyses (NMA) are essential for surgical research, a domain that is classically limited by small sample sizes with many barriers to completing high-quality large-scale trials [1, 2]. While traditional meta-analyses are limited to direct comparison of two interventions, NMA provides a means to perform both direct and indirect comparisons across multiple interventions of interest [3]. It is a statistical technique used in evidence synthesis to compare multiple interventions or treatments through direct and indirect comparisons among existing primary studies [4, 5]. By combining studies that link three or more interventions, you are forming a network of interventions, hence the name, “network meta-analyses.”

L. J. Park (✉)
Division of General Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada

Department of Health Research Methods,
Evidence, and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: lily.park@medportal.ca

K. Um
Department of Health Research Methods,
Evidence, and Impact, McMaster University,
Hamilton, ON, Canada

Department of Medicine, Division of Cardiology,
McMaster University, Hamilton, ON, Canada
e-mail: kevin.um@medportal.ca

D. Zera
Department of Health Research Methods,
Evidence, and Impact, McMaster University,
Hamilton, ON, Canada

E. Duceppe
Department of Medicine, Department of Internal
Medicine, CHUM, Montreal, ON, Canada
e-mail: Emmanuelle.duceppe@phri.ca

P. E. Serrano
Division of General Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence & Impact, McMaster University, Hamilton,
ON, Canada
e-mail: serrano@mcmaster.ca

Table 32.1 Comparison of network meta-analyses compared to traditional pairwise meta-analyses

	Network meta-analysis	Pairwise meta-analysis
Number of comparators	>2	2
Types of studies included	Comparative studies, ideally randomized controlled trials	Comparative studies, ideally randomized controlled trials
Evidence synthesis	Can synthesize both direct and indirect evidence	Can only synthesize direct evidence
Model outputs	Summary effect estimate with relative effects to create hierarchy of interventions	Summary effect only

The main elements of NMA that sets it apart from other methods of evidence synthesis are as follows and demonstrated in Table 32.1 [3–7]:

- Compare three or more interventions in a single analysis
- Combine direct and indirect evidence (i.e., you can compare two interventions indirectly, when no studies have compared them directly)
- Estimate relative effects between any pair of interventions in the network you created, which allows for greater precision of estimates and estimation of ranking to create a hierarchy of interventions

Despite its utility and high-level status in the hierarchy of evidence, the complexity of the concepts and statistical analyses that underscore NMAs can render interpretation to be a challenge. This chapter will help guide the reader through understanding and critically appraising NMAs in consideration for use in clinical practice.

Clinical Scenario

You are hepatobiliary surgeon seeing Mr. X for his preoperative appointment. Mr. X. is a 68-year-old male with hypertension, dyslipidemia, and non-insulin dependent Type 2 diabetes mellitus. He had a known rectal tumor with metastatic disease to the liver, which you have deemed to be technically resectable. He has already undergone resection of his primary rectal cancer. His recovery from this surgery was complicated by a surgical site infection (SSI). He asks what steps you’ll

be taking to avoid another SSI. You discuss the perioperative sterility practice including preparatory solution prior to the first incision. The patient is curious as to what solution you will be using and asks for you to use the “strongest antiseptic you have.” You have always used a 2.0% chlorhexidine solution but are unsure of any updated literature that suggest other options that may confer decreased SSI risk. You decide to search the literature for evidence of optimal preoperative preparatory solutions for this clean-contaminated surgery you plan to book.

Literature Search

You begin your literature search through a clinical question framed using the PICO(T) format that includes specification about the patient population (P), intervention (I), comparative interventions (C), outcomes (O), and time period (T) (see Chap. 3). The question is posed as follows: In patients undergoing elective clean-contaminated surgeries, is 2.0% chlorhexidine, more effective compared to other preparation solutions, at reducing the risk of SSI?

Using Medline, the National Library of Medicine’s PubMed database [8], you perform a literature review by entering “surgical site infection” AND “skin antiseptic” AND “preoperative” in the search field. You limit your search to English articles, published in the last 3 years, meta-analysis, and human subjects (Appendix). The search results in eight articles [9–16]. Among these reviews, three are network meta-analyses on the topic of interest [10, 11, 14]. You decide to review the study that

considered wound classifications of participants and included the largest number of participants, which is entitled, “Efficacy of different preoperative skin antiseptics on the incidence of surgical site infections: a systematic review, GRADE assessment, and network meta-analysis” by Jalalzadeh et al. that was published in October 2022 [10].

Summary of the Appraised Article

This article, published in *The Lancet Microbe*, is a systematic review and network meta-analysis including 27 studies with 17,735 patients that compared different preoperative skin antiseptics in the prevention of SSIs in adult patients undergoing surgery of any wound classification [10]. This review included only randomized controlled trials (RCT) that directly compared two or more antiseptic agents (namely, chlorhexidine, iodine, or olanexidine) or different concentrations of antiseptics in aqueous and alcohol-based solutions [10].

Results of the study demonstrate that compared to aqueous iodine, 2.0–2.5% chlorhexidine in alcohol (relative risk 0.75, 95% CI 0.61–0.92) and 1.5% olanexidine (RR 0.49, 95% CI 0.26–0.92) significantly reduced the rate of SSIs [10]. Their conclusion was that for adult patients undergoing a surgical procedure of any wound classification, skin preparation using either 2.0–2.5% chlorhexidine in alcohol or 1.5% olanexidine is most effective in the prevention of SSIs. For clean surgery, no specific concentration of chlorhexidine in alcohol can be recommended.

To determine appropriateness of applying these results to your practice, it is essential to understand the quality of this evidence. Therefore, you start at the basic concepts underpinning NMAs.

Evaluating a Network Meta-analysis

It is important for the clinician to always assess whether a study was conducted with robust methodologies to produce reliable

results. Of note, three main questions should be asked: Are the results valid? What are the results? And are the results applicable to my practice? The important points to consider when evaluating a network meta-analysis are outlined in Box 32.1. The list was modified from [17].

Box 32.1: Important Points to Consider when Evaluating a Network Meta-analysis

Are the results valid?

- Did the review explicitly address a sensible clinical question?
- Was the search for relevant studies detailed and exhaustive?
- Were the primary studies of high methodologic quality?
- Were assessments of studies reproducible?

What are the results?

- Were the results similar from study to study?
- What are the overall results of the review?
- How precise were the results?

How can I apply the results to patient care?

- How can I best interpret the results to apply them to the care of my patients?
- Were all clinically important outcomes considered?
- Are the benefits worth the costs and potential risks?

These are the questions we will keep in mind and answer as we navigate assessment of the study by Jalalzadeh et al. [10]

Fundamentals of Network Meta-analysis

Inclusion Criteria for Studies in NMAs

NMAs, like traditional meta-analyses, can include comparative studies that are not necessarily RCTs (see Table 32.1). The Jalalzadeh et al. study limited inclusion of primary studies to

RCTs [10]. This is especially prudent in NMAs as indirect comparisons across RCTs are considered observational evidence and therefore are at risk of observational study biases, importantly confounding [3]. Therefore, to draw indirect comparisons between primary observational comparative studies would exacerbate these biases.

Transitivity is an important concept in NMA that refers to the principle that all RCTs included in NMAs should be similar in almost all characteristics that can impact the effect of the intervention [3–5, 18]. In essence, transitivity requires that all interventions that will be included in the network of evidence are similar across all the included studies with regard to characteristics that may affect the treatment effects [19, 20]. These characteristics would be considered effect modifiers [19, 20]. For instance, if there are two trials, one (1) that compares preparatory solution A and B and another (2) that compares preparatory solution A and C, A should be similar in how it is utilized in both studies. If study 1 investigates the utility of solution A versus B applied twice at 10-min intervals before the first incision meanwhile study 2 investigates A versus C following a single application, the inclusion of both for synthesis in an NMA would violate the principle of transitivity. The double application of solutions in study 1 represents an effect modifier—it is a study characteristic that may impact the effect of the intervention. If ever there is

doubt about whether the inclusion of a trial may violate transitivity, remind yourself that transitivity requires all interventions included in the NMA to be jointly randomizable [3, 21, 22]. To that point, if you’re uncertain about whether a study can be included in your NMA, you should ask yourself: would you include all interventions across the individual primary studies in one single multi-arm randomized trial? If your answer is “no,” you may be at risk of defying transitivity.

Understanding Comparisons in NMA

A given NMA may provide various treatment comparisons including indirect, direct, or a mix of both. A simple way to understand the comparisons that occur in a NMA is to imagine a network of nodes and edges (Fig. 32.1). Consider you are trying to make comparisons between four interventions A, B, C, and D. If you have studies that compare these four interventions (i.e., AB, AC, CD), you can conduct indirect comparisons between treatments that share a “connection” in comparison, which does not necessarily have to share a single common comparator. For instance, since AB and AC share a comparison “connection” through treatment A, you are able to indirectly compare B and C. Similarly, since AC and CD share a comparison “connection” through treatment C, you are able to indirectly compare A and D. However, B and D do not share a single

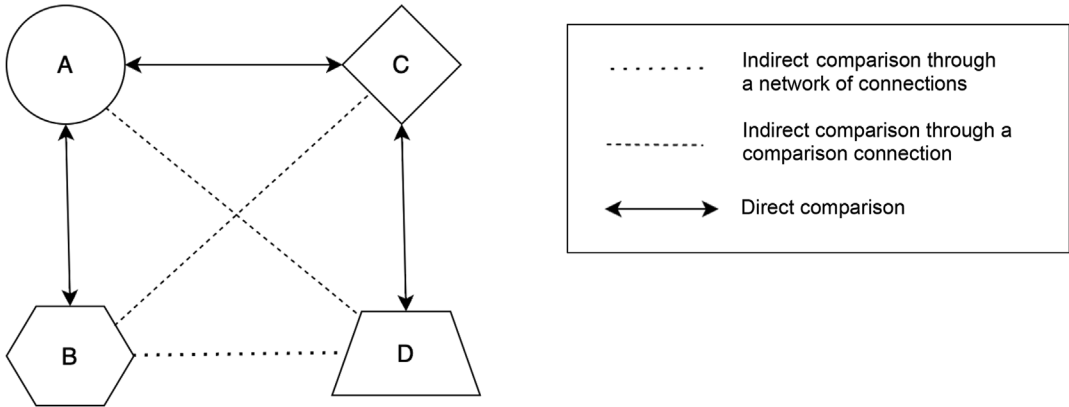


Fig. 32.1 Possible indirect comparisons that can be made through network meta-analyses of existing studies with direct analyses

common comparison “connection” in either A or C. Still, they can be compared indirectly through the network connection of comparisons.

Are the Results Valid?

Did the Review Explicitly Address a Sensible Clinical Question?

Firstly, we want to assess that the article addressed a sensible clinical question and appropriately included studies relevant to that question. The Jalalzadeh et al. NMA included only RCTs that compared two or more antiseptic skin preparation agents or concentrations in aqueous and alcohol-based solutions in adults undergoing surgical procedures in the operating theater that reported SSI rates. This is consistent with the interventions and outcome of interest. It is important to note that this study included any surgical procedures that occurred in the operating room, while our patient will be undergoing an intraabdominal, clean-contaminated surgery. However, they did perform subgroup analyses by wound classification (i.e., clean, clean-contaminated, contaminated, or dirty), which was preplanned [10]. In this way, the review explicitly answered a sensible clinical question.

Was the Search for Relevant Studies Detailed and Exhaustive?

The authors report a thorough search methodology for studies that meet their inclusion criteria. They explicitly state the use of bibliographic databases, specifically Embase, PubMed, and Cochrane CENTRAL. They also identified additional articles through backward and forward citation tracking of published reviews and included studies. Their complete search criteria are also provided in the supplementary. Altogether, they have conducted a detailed and exhaustive search.

Were the Primary Studies of High Methodologic Quality?

Looking at Appendix 8 in the supplementary material for the Jalalzadeh et al. article, most studies included in the review were deemed to have some concerns for bias. There were two studies with low risk of bias and seven studies that were deemed to be high risk of bias. However, Jalalzadeh et al. conducted sensitivity analyses excluding the studies that were deemed to be overall high risk of bias [10]. Still, the results of this review should be considered in the context of most studies having some concerns for bias.

Were Assessments of Studies Reproducible?

The authors report that screening for studies were conducted independently and in duplicate with consultation of a senior author for disagreements, which were resolved by consensus. Intra-class correlations were not reported, which limits our ability to comment on reproducibility of study assessments.

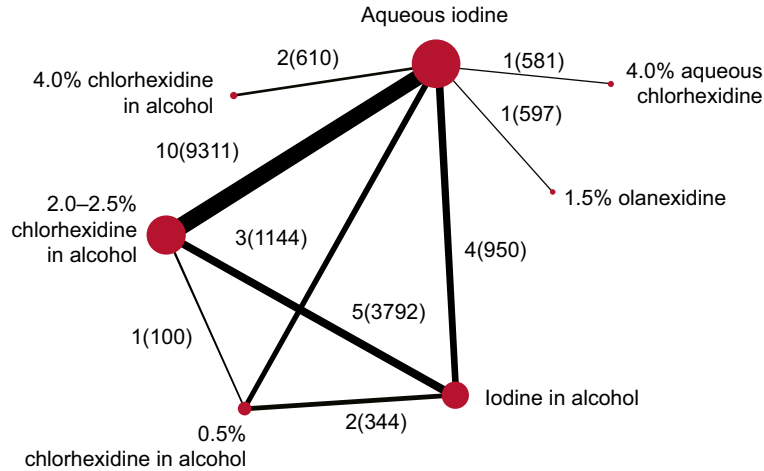
What Are the Results?

What Are the Overall Results of the Review?

Network Graphs

Network graphs provide a visual representation of the comparisons present among the primary studies included in the NMA. It is comprised of dots and lines with the largest dots demonstrating the treatment that is used most frequently as a comparator across included studies, and the thickest line representing the comparison with the largest number of studies. Sometimes colors can also be used to demonstrate the risk of bias for the primary comparisons [4, 5].

Fig. 32.2 Network graph of interventions from study by Jalalzadeh et al. [10]



In the selected article by Jalalzadeh et al., the authors provide a network graph, which clearly depicts aqueous iodine to be the most frequently studied intervention followed by 2.0–2.5% chlorhexidine in alcohol, then iodine in alcohol (Fig. 32.2) [10]. It also demonstrates that the most frequent comparison was made between the former two solutions. This provides a great overview on all the interventions that have been studied in the context of interest while also providing a visual distribution of comparisons that have been conducted.

Statistical Analyses

Now that we have established the network of evidence defined by the authors, we can consider the statistical analyses that were undertaken. Before diving into the network, the authors appropriately conducted pairwise meta-analyses with studies that directly compared two treatments within the network. From there, the authors developed a network model that considers both the direct and indirect evidence through common comparators to estimate treatment effects [3, 22, 23].

There are different statistical frameworks under which, a NMA can be conducted. These include a Bayesian or frequentist approach [3, 24]. Bayesian approach takes into consideration existing or prior beliefs about the relationships between treatments [3, 24]. Using Bayes’ theorem, the collected data and analyses would help update the prior beliefs [3, 24]. Bayesian analy-

ses provide a range of possible values within which the true treatment effect is likely to fall. On the other hand, the Frequentist approach solely focuses on the data at hand in the present study as opposed to incorporating prior beliefs. The frequentist approach would provide point estimates and confidence intervals based on the analyzed data [3, 24]. The selection between the Bayesian or frequentist approach would therefore depend on the nature of the research question, the availability of prior information, and the preferences of the researcher [3, 24].

It is also important to determine whether a fixed-effect model or a random-effects model will be used. A fixed effect model assumes that all studies included in the analysis have the same underlying or “fixed” treatment effect such that any observed differences are due to random sampling error [25]. Random effect models acknowledge that the true intervention effect may vary between studies, and therefore account for both within-study and between-study variability [25].

In the case of the Jalalzadeh et al. study, they planned to perform the NMA using the frequentist method and random effects models [10].

League Tables and Ranking Metrics

League tables also provide a visual representation of all comparisons but with a focus on relative treatment effects. The standard format encompasses all interventions included in the NMA down the diagonal axis of the table. All the

effect estimates from the NMA will be presented below the diagonal list of interventions, while all the direct effect estimates reported from individual studies will be presented above the diagonal. Since NMAs generate effect estimates from indirect comparisons, it can be normal, and, in fact, expected to have some missing entries above the diagonal, while all entries below the diagonal are completed with the NMA estimates. Additionally, there may be discrepancies in the effect estimates for the comparison above and below the diagonal due to the network estimates taking into account all direct and indirect evidence while the pairwise estimates (i.e., estimates above the diagonal) were calculated using only direct evidence.

Table 32.2 demonstrates the League table from the Jalalzadeh et al. study, which provides all pairwise relative treatment effects [10]. Here, they have bolded statistically significant differ-

ences. For instance, they found that 2.0–2.5% chlorhexidine in alcohol has a relative risk of 0.75 (0.61–0.92) compared to aqueous iodine in developing SSIs [10].

League tables are often accompanied by a metric that ranks all the interventions included in the study. For NMA conducted under a Bayesian framework, treatment ranking would generally be conducted using the surface under the cumulative ranking curve (SUCRA) [26]. The SUCRA analogue when using the frequentist method are netranking p-scores [26, 27].

Jalalzadeh et al. provides Netranking P-scores, as they used the frequentist method for their data analysis [10]. The P-score is a ranking metric for NMA that varies from 0 to 1. A treatment with a higher score is considered to be more effective than a treatment with a lower score [10, 26, 27]. These scores are based on point estimates and

Table 32.2 League table from Jalalzadeh et al. study

A. All wound classifications/any type of surgery						
<i>League table</i>						
Olanexidine 1.5%	.	.	.	.	.	0.49 (0.26–0.92)
0.73 (0.27–1.92)	Alcoholic CHG 4.0%	.	.	.	.	0.67 (0.32–1.40)
0.70 (0.33–1.48)	0.97 (0.42–2.23)	Alcoholic CHG 0.5%	3.00 (0.61–14.74)	.	0.33 (0.08–1.44)	0.68 (0.45–1.03)
0.65 (0.33–1.27)	0.90 (0.42–1.93)	0.93 (0.60–1.43)	Alcoholic CHG 2.0–2.5%	.	0.76 (0.57–1.03)	0.77 (0.62–0.97)
0.52 (0.15–1.42)	0.72 (–.25–2.10)	0.75 (0.31–1.78)	0.80 (–.36–1.79)	Aqueous CHG 4.0%	.	0.93 (0.42–2.01)
0.50 (0.25–1.01)	0.69 (0.31–1.53)	0.71 (0.45–1.14)	0.77 (0.60–1.00)	0.96 (0.42–2.19)	Alcoholic iodine	0.87 (0.54–1.40)
0.49 (0.26–0.92)	0.67 (0.32–1.40)	0.69 (0.47–1.02)	0.75 (0.61–0.92)	0.93 (0.43–2.01)	0.97 (0.73–1.29)	Aqueous iodine
<i>Netranking P-score</i>						
Olanexidine		0.8866				
Alcoholic CHG 0.5%		0.6517				
Alcoholic CHG 4.0%		0.6345				
Alcoholic CHG 2.0–2.5%		0.5891				
Aqueous CHG 4.0%		0.3406				
Alcoholic iodine		0.2257				
Aqueous iodine		0.1719				

standard errors of the frequentist NMA estimates, which quantify the probability that a treatment is better than the competing treatments [10, 26, 27]. Of note, they do not incorporate assessments of the quality of evidence. Thus, the ranking should always be interpreted with caution, with careful consideration for the confidence in the results [3, 28]. Other ranking metrics include the display of probabilities that each comparator could take a certain rank, either in a table otherwise known as “bubble-plots” or rankograms (numerical summary of rankograms are provided by SUCRA) [26, 28, 29]. It should also be noted that ranking metrics from NMAs are often under criticism for oversimplifying a complex set of information into a ranking system that is subject to instability and misinterpretation [30–33]. Considering the controversies surrounding treatment ranking from NMAs, we should refrain from making clinical decisions based on reported ranking metrics alone. Instead, the focus should be on the relative treatment effects and their uncertainty, which is generally agreed to be the most clinically relevant and reliable output from NMAs [22, 34, 35].

Forest Plots

Forest plots provide pairwise meta-analyses between one common comparator and the remaining comparators. Figure 32.3 is a forest plot from Jalalzadeh et al. that demonstrates the effect estimates and corresponding 95% confidence intervals in the pairwise analyses of various solutions compared to aqueous iodine (common comparator) [10]. Usually, these forest

plots choose the common comparator that is most commonly compared in direct analyses, which in this case was aqueous iodine.

Were the Results Similar from Study to Study?

Effect estimates reported by individual studies in the Jalalzadeh et al. review, are not explicitly reported in the study. In Appendix 3 of the supplementary, the authors provide the total number of SSI events per study, but not rates by the different solution groups. However, we can infer whether the results were similar between studies by looking at the effect estimates presented above the diagonal of comparators in the league table. Effect estimates with narrower confidence intervals imply that there were more similar results between individual studies. However, the best way to assess would be for the reader to look at each individual study and their reported effect estimates.

How Precise Were the Results?

Statistically significant results of the article include the superiority of 2.0–2.5% chlorhexidine in alcohol and 1.5% olanexidine compared to aqueous iodine. The relative risk estimates were 0.75 (95% CI 0.61–0.92) and 0.49 (95% CI 0.26–0.92), respectively. The forest plot provides a better depiction of the estimate precision. The smaller confidence interval for the comparison

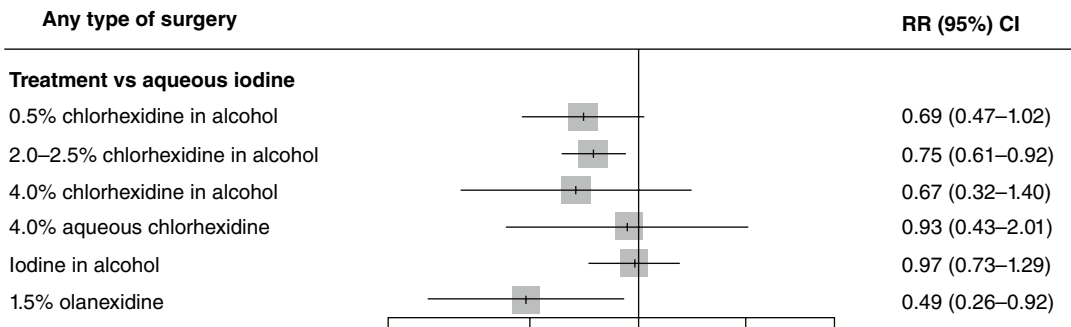


Fig. 32.3 Forest plot from Jalalzadeh et al. study

with 2.0–2.5% chlorhexidine in alcohol suggests greater precision compared to the larger confidence interval for the comparison with 1.5% olanexidine. From the network graph, we were able to see that aqueous iodine compared to 2.0–2.5% chlorhexidine in alcohol was the most frequently studied comparison, which likely contributed to greater precision of the estimates compared to that of 1.5% olanexidine.

Consistency or Coherence Checks

Now that we understand the main message of the study, we would like to determine our confidence in their results. We previously discussed that transitivity is the core assumption that underpins the validity of NMAs. Thus, it is crucial that all NMAs assess for transitivity.

Qualitative Check for Consistency/Coherence

The statistical analogue of transitivity is coherence, which is also referred to as consistency [3, 34, 36]. Coherence is the assumption that all the sources of evidence, both direct and indirect, agree with each other in terms of the intervention effects for a specific comparison [3]. Given this definition, we cannot have coherence without transitivity and vice versa. Therefore, it is crucial that coherence checks are conducted throughout the development of an NMA.

We previously discussed the strategy of asking yourself if interventions from all included studies would be jointly randomizable at the level of study inclusion. That is, confirming that we could reasonably perform a “mega-trial” with all comparators in the setting and manner in which they were delivered in their respective trials.

Upon review of the individual studies, the distribution of potential effect modifiers should also be considered [3, 34, 36]. For instance, if we were to believe that the severity of obesity may impact intervention effect of preparation solutions on SSI development, we would need to evaluate the distribution of body mass indices

across participants enrolled in the trials that we included. If an RCT comparing solution A versus B (study I) was conducted across obese patients only while an RCT comparing solution B versus C (study II) was conducted across non-obese patients only, the distribution of effect modifiers would be too different to allow for a valid indirect effect estimate. In reality there may be variability in the distribution of obese patients; for instance, study I can have 30% obese participants while study II has 80% obese participants. In these situations, it would be important to determine if obesity is an important enough effect modifier, to render these indirect estimates invalid considering the differences in the distribution of obesity between study I and II.

Quantitative Check for Consistency/Coherence

A more quantifiable check for coherence can and should also be conducted through evaluation of an incoherence factor or inconsistency factor (IF) [3, 37]. The IF measures disagreement between the direct and indirect effect estimates for any pairwise comparisons in the network. This test for incoherence can be conducted at local and global levels [3, 37]. Local incoherence tests evaluate for specific areas of the evidence network that may be subject to incoherency while global incoherence tests evaluate the coherence of the entire evidence network as a whole [3, 37].

Evaluation for Local Incoherence

Checking for local incoherence can be done through the Separating Indirect from Direct Evidence (SIDE) approach, which is also referred to as “node-splitting” [38]. This method evaluates the IF for all comparisons. As the name implies, node-splitting “splits” the evidence for a specific comparison into direct and indirect components [38]. Then, a statistical test is performed to assess whether the direct and indirect evidence are consistent with each other or if there is a significant discrepancy [38]. Generally, a p-value

<0.05 is considered to be statistically significant, suggesting detection of incoherence in the identified region of the evidence network [38].

Evaluation for Global Incoherence

There are different methods to evaluate global incoherence including incoherence models where IFs across the entire network are formally tested with a χ^2 test [3, 37]. More commonly, the Q-test and I^2 test that are used to test for heterogeneity in pairwise analyses have been adapted to evaluate for heterogeneity and incoherence in NMAs [3, 37].

The Jalalzadeh et al. study conducted the quantitative coherence checks to assess for transitivity of the treatments [10]. The node-splitting method did not demonstrate statistically significant inconsistencies between direct and indirect comparisons, and heterogeneity between studies was assessed to be moderate, as the I^2 value was 27.5% [10]. Although these metrics are reassuring, the reader should remember that conclusions cannot be drawn solely from the presence or absence of statistical significance [3, 37]. Due to the low power of tests for incoherence, they may fail to detect incoherence even when they are present [39, 40]. Therefore, it is important to conduct the qualitative coherence checks as well as the quantitative checks to ensure transitivity of the network and validity of the NMA effect estimates [3, 37, 39, 40].

Evaluating the Confidence in the Evidence

The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework is widely used to evaluate the quality of evidence and strength of recommendations in systematic reviews and practice guidelines [41, 42]. The five main domains of the standard GRADE system have also been modified to fit network meta-analyses [3]. Assuming the reader has a general understanding of the GRADE framework, we will focus on the modifications

for NMAs and how they have been applied to the Jalalzadeh et al. article [10].

Study Limitations

The decision to downgrade based on classical risk of bias items for all individual studies can be influenced in different ways. Salanti and colleagues have suggested for authors to generate a bar graph that demonstrates different confidence ratings of existing direct estimates [3]. Interpretation of this graph can help determine whether our confidence in the evidence needs to be downgraded. Most studies included in the Jalalzadeh et al. article were considered to have some concerns for bias with seven studies deemed to be high risk of bias. A network sensitivity meta-analysis was conducted excluding studies with high risk of bias. This demonstrated comparable results. Therefore, the authors did not downgrade for risk of bias.

Indirectness

Indirectness in NMAs calls upon the assessment of transitivity and similarity of the target research question (i.e., PICO[T] elements) among the included studies. The authors assessed for similarity of studies throughout the process of primary study inclusion and determined that transitivity was not violated through their qualitative checks [10].

Inconsistency

Inconsistency refers to the assessment of heterogeneity within direct comparisons and tests for incoherence. It is recommended to consider these two types of variation and assess them jointly to determine whether downgrading for inconsistency is necessary. In the Jalalzadeh et al. study, heterogeneity across studies was moderate with an I^2 value of 27.5%. Furthermore, node-splitting did not demonstrate inconsistencies between direct and indirect comparisons and so the results were not downgrade for inconsistency [10].

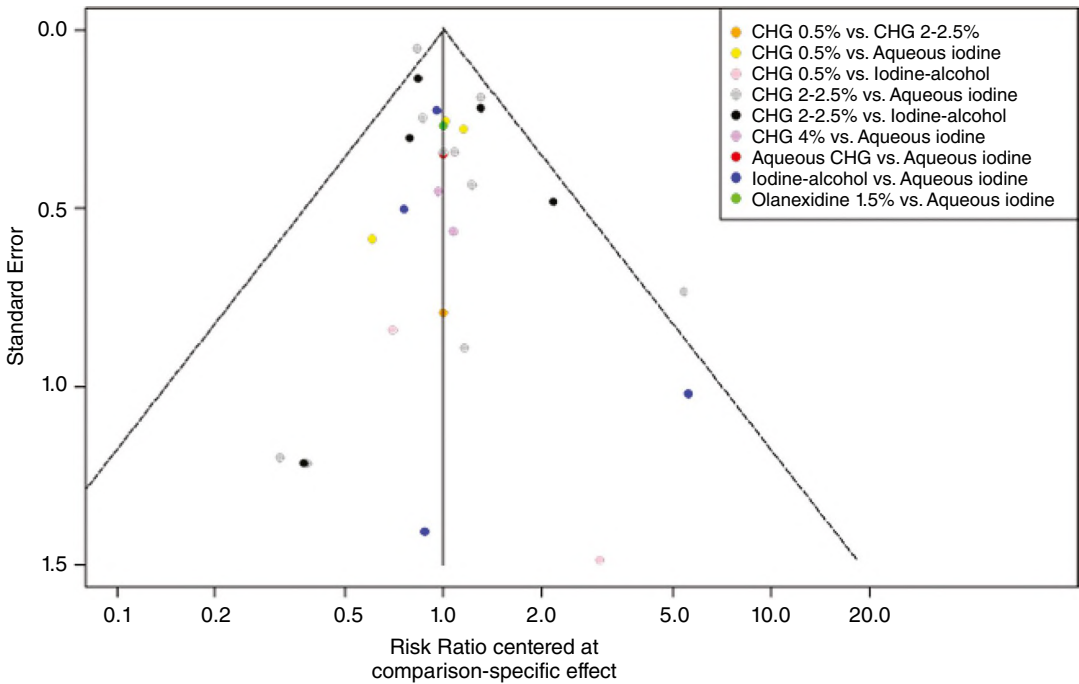


Fig. 32.4 Comparison-adjusted funnel plot created by Jalalzadeh et al. [10] to assess for publication bias

Imprecision

Imprecision, similar to conventional pairwise meta-analyses, can be judged based on their 95% confidence intervals [3]. In the Jalalzadeh et al. article, relative risk estimates with CI that crossed the null effect threshold ($RR = 1$) was downgraded by 1 [10]. The exception was the olanexidine 1.5% versus aqueous iodine comparison [10]. Although there was a large effect and 95% CI did not cross the threshold of null effect, the optimal information size was not met, as only one RCT was available to directly compare these two solutions [10]. Therefore, this result was downgraded by 1 for imprecision [10].

Publication Bias

Publication bias can be assessed in NMAs by using a comparison-adjusted funnel plot, which demonstrates the effect estimate of a study versus its precision for SSI [3]. This helps assess for small-study effects in NMA [3, 43]. The funnel plot generated by Jalalzadeh et al. (Fig. 32.4)

shows symmetry and therefore indicates that there are likely no differences between small and large studies with regard to the effect of treatment (i.e., no small-study effect) [10]. Asymmetry in comparison-adjusted funnel plots can also be caused by publication bias, and so here, publication bias is not likely [3, 34]. However, readers should also consider whether there could be unpublished studies for all pairwise comparisons in the network that could contribute to bias of the results [3, 43].

How Can I Apply the Results to Patient Care?

How Can I Best Interpret the Results to Apply Them to the Care of My Patients?

Summarizing Results

At this point in our appraisal of the study by Jalalzadeh et al., we have determined that the authors identified direct evidence and have established a network of evidence. The NMA deter-

mined that compared to iodine, aqueous iodine 2–2.5% chlorhexidine in alcohol and 1.5% olanexidine significantly reduced the rate of SSIs with relative risks of 0.75 (95% CI 0.61–0.92) and 0.49 (95% CI 0.26–0.92), respectively [10].

This was based on a review that included 33 RCTs, of which 27 RCTs with 17735 participants were included in the NMA. The entire network of evidence included 27 different solutions that were categorized into 7 broad groups according to the type and concentration of the solution.

Although there is benefit with 1.5% olanexidine compared to iodine for prevention of SSIs in clean-contaminated surgery, this is considered moderate certainty of evidence since the effect was only demonstrated in a single randomized trial. Different concentrations of chlorhexidine suggest benefit compared to iodine, but the effect estimates remain statistically insignificant, contributed by the relatively small number of studies and wide confidence intervals.

Therefore, given the existing evidence, 2.0–2.5% chlorhexidine in alcohol remains the best option for the prevention of SSI in adult patients undergoing a surgical procedure of any wound classification.

Were All Clinically Important Outcomes Considered?

The primary outcome of this review were SSIs. However, the review also investigated adverse events of the skin related to the different skin preparation solutions. Considering the context and complexity of NMAs, the focus on SSI as the primary outcome with an assessment of its safety was adequate in understanding the most clinically important outcomes as they relate to our initial clinical question.

Are the Benefits Worth the Costs and Potential Risks?

Costs were not reported in the study; however, they did investigate for adverse events of the skin to different preparation solutions. The

review reports that among the ten RCTs that mentioned adverse events, there were no differences among the skin preparation groups. Therefore, the benefits in selecting 2.0–2.5% chlorhexidine in alcohol likely outweigh any potential risks.

Return to Case Scenario

You see Mr. X again on the morning of his surgery to re-discuss the antiseptic solution. You discuss that the current evidence does not demonstrate benefit with higher concentrations of chlorhexidine in reducing SSI. Although there is a newer agent, 1.5% olanexidine, that does demonstrate efficacy in protection against SSI, there is not enough data to strongly recommend this over the usual 2.0–2.5% chlorhexidine in alcohol at this time. Mr. X is satisfied with this explanation and feels comfortable to proceed with his surgery.

Key Points for Clinicians

- Determine that the study you are reviewing is answering the appropriate research question relevant to your clinical scenario
- When assessing an NMA, ensure that transitivity assumption is not violated by looking for whether the authors conducted qualitative and quantitative coherency checks
- Look for network graphs to gain a visual understanding of the available evidence and the network created by the NMA authors
- Ranking metrics in NMAs must be interpreted with caution
- Consider the adjustments in applying the GRADE framework to evaluating the certainty of evidence from NMAs

Appendix: Literature Search Strategy

“surgical site infection” AND “skin antiseptic” AND “preoperative”

Limited to:

- Article type: Meta-analysis
- Publication date: 2021-01-01 to 2024-01-01,
- Article language: English
- Species: Human

References

1. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. *BMJ*. 2002;324(7351):1448–51.
2. Robinson NB, Fremes S, Hameed I, Rahouma M, Weidenmann V, Demetres M, Morsi M, Soletti G, Di Franco A, Zenati MA, Raja SG, Moher D, Bakaeen F, Chikwe J, Bhatt DL, Kurlansky P, Girardi LN, Gaudino M. Characteristics of randomized clinical trials in surgery from 2008 to 2020. *JAMA Netw Open*. 2021;4(6):e2114494.
3. Chaimani A, Caldwell DM, Li T, Higgins JPT, Salanti G. Chapter 11: Undertaking network meta-analyses [Internet]. [cited 2023 Dec 17]. Available from: <https://training.cochrane.org/handbook/current/chapter-11>
4. Rouse B, Chaimani A, Li T. Network meta-analysis: an introduction for clinicians. *Intern Emerg Med*. 2017;12(1):103–11.
5. Phillips MR, Steel DH, Wykoff CC, Busse JW, Bannuru RR, Thabane L, Bhandari M, Chaudhary V. A clinician's guide to network meta-analysis. *Eye*. 2022;36(8):1523–6.
6. Cooper NJ, Peters J, Lai MCW, Juni P, Wandel S, Palmer S, Paulden M, Conti S, Welton NJ, Abrams KR, Bujkiewicz S, Spiegelhalter D, Sutton AJ. How valuable are multiple treatment comparison methods in evidence-based health-care evaluation? *Value Health*. 2011;14(2):371–80.
7. Caldwell DM, Dias S, Welton NJ. Extending treatment networks in health technology assessment: how far should we go? *Value Health*. 2015;18(5):673–81.
8. Motschall E, Falck-Ytter Y. Searching the MEDLINE literature database through PubMed: a short guide. *Oncol Res Treat*. 2005;28(10):517–22.
9. Li L, Wang Y, Wang S. Efficacy comparison of chlorhexidine and iodine preparation in reduction of surgical site infection: a systemic review and meta-analysis. *Int J Nurs Stud*. 2022;127:104059.
10. Jalalzadeh H, Groenen H, Buis DR, Dreissen YE, Goosen JH, Ijpma FF, van der Laan MJ, Schaad RR, Segers P, van der Zwet WC, Griekspoor M, Harmsen WJ, Wolfhagen N, Boormeester MA. Efficacy of different preoperative skin antiseptics on the incidence of surgical site infections: a systematic review, GRADE assessment, and network meta-analysis. *Lancet Microbe*. 2022;3(10):e762–71.
11. Wade RG, Burr NE, McCauley G, Bourke G, Efthimiou O. The comparative efficacy of chlorhexidine gluconate and povidone-iodine antiseptics for the prevention of infection in clean surgery: a systematic review and network meta-analysis. *Ann Surg*. 2021;274(6):e481–8.
12. Aftab R, Dodhia VH, Jeanes C, Wade RG. Bacterial sensitivity to chlorhexidine and povidone-iodine antiseptics over time: a systematic review and meta-analysis of human-derived data. *Sci Rep*. 2023;13(1):347.
13. Chen Z, Mont MA. The utility of chlorhexidine cloth use for the prevention of surgical site infections in total hip arthroplasty and surgical as well as basic science applications: a meta-analysis and systematic review. *Orthop Clin North Am*. 2023;54(1):7–22.
14. Hampton J, Park SSW, Palazzi K, Oldmeadow C, Carroll R, Attia J, Smith SR. The effect of preoperative skin preparation on clinical outcomes with incisional surgery: a network meta-analysis. *ANZ J Surg*. 2022;92(11):2859–67.
15. Mastrocola M, Matziolis G, Böhle S, Lindemann C, Schlattmann P, Ejler H. Meta-analysis of the efficacy of preoperative skin preparation with alcoholic chlorhexidine compared to povidone iodine in orthopedic surgery. *Sci Rep*. 2021;11(1):18634.
16. Peristeri DV, Nour HM, Ahsan A, Abogabal S, Singh KK, Sajid MS. Alcohol-containing versus aqueous-based solutions for skin preparation in abdominal surgery: a systematic review and meta-analysis. *J Surg Res*. 2023;291:734–41.
17. Bhandari M, Devereaux PJ, Montori V, Cinà C, Tandan V, Guyatt GH. Users' guide to the surgical literature: how to use a systematic literature review and meta-analysis. *Can J Surg*. 2004;47(1):60–7.
18. Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JPT. Evaluating the quality of evidence from a network meta-analysis. *PLoS One*. 2014;9(7):e99682.
19. Song N, Gao M, Tian J, Gao Y. Choice for prophylactic antibiotics: Cefazolin plus metronidazole or cefoxitin alone for endometrial cancer surgical staging. *J Obstet Gynaecol Res*. 2020;46(9):1864–70.
20. Glenny AM, Altman DG, Song F, Sakarovich C, Deeks JJ, D'Amico R, Bradburn M, Eastwood AJ, International Stroke Trial Collaborative Group. Indirect comparisons of competing interventions. *Health Technol Assess*. 2005;9(26):1–134, iii–iv.
21. Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ*. 2005;331(7521):897–900.
22. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Res Synth Methods*. 2012;3(2):80–97.
23. Seide SE, Jensen K, Kieser M. A comparison of Bayesian and frequentist methods in random-effects network meta-analysis of binary data. *Res Synth Methods*. 2020;11(3):363–78.
24. Sadeghirad B, Foroutan F, Zoratti MJ, Busse JW, Brignardello-Peterson R, Guyatt G, Thabane L. Theory and practice of Bayesian and frequentist

- frameworks for network meta-analysis. *BMJ Evid Based Med.* 2023;28(3):204–9.
25. Dettori JR, Norvell DC, Chapman JR. Fixed-Effect vs random-effects models for meta-analysis: 3 points to consider. *Global Spine J.* 2022;12(7):1624–6.
 26. Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Med Res Methodol.* 2015;15(1):58.
 27. Rosenberger KJ, Duan R, Chen Y, Lin L. Predictive P-score for treatment ranking in Bayesian network meta-analysis. *BMC Med Res Methodol.* 2021;21(1):213.
 28. Salanti G, Nikolakopoulou A, Efthimiou O, Mavridis D, Egger M, White IR. Introducing the treatment hierarchy question in network meta-analysis. *Am J Epidemiol.* 2022;191(5):930–8.
 29. Mbuagbaw L, Rochwerf B, Jaeschke R, Heels-Andsell D, Alhazzani W, Thabane L, Guyatt GH. Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev.* 2017;6(1):79.
 30. Trinquart L, Attiche N, Bafeta A, Porcher R, Ravaud P. Uncertainty in treatment rankings: reanalysis of network meta-analyses of randomized trials. *Ann Intern Med.* 2016;164(10):666–73.
 31. Veroniki AA, Straus SE, Rücker G, Tricco AC. Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? *J Clin Epidemiol.* 2018;100:122–9.
 32. Mills EJ, Kanters S, Thorlund K, Chaimani A, Veroniki AA, Ioannidis JPA. The effects of excluding treatments from network meta-analyses: survey. *BMJ.* 2013;347:f5195.
 33. Kibret T, Richer D, Beyene J. Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study. *Clin Epidemiol.* 2014;6:451–60.
 34. Cipriani A, Higgins JPT, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. *Ann Intern Med.* 2013;159(2):130–7.
 35. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, Ioannidis JPA, Straus S, Thorlund K, Jansen JP, Mulrow C, Catalá-López F, Gøtzsche PC, Dickersin K, Boutron I, Altman DG, Moher D. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Ann Intern Med.* 2015;162(11):777–84.
 36. Jansen JP, Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. *BMC Med.* 2013;11(1):159.
 37. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Stat Assoc.* 2006;101(474):447–59.
 38. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med.* 2010;29(7–8):932–44.
 39. Veroniki AA, Mavridis D, Higgins JP, Salanti G. Characteristics of a loop of evidence that affect detection and estimation of inconsistency: a simulation study. *BMC Med Res Methodol.* 2014;14(1):106.
 40. Song F, Clark A, Bachmann MO, Maas J. Simulation evaluation of statistical properties of methods for indirect and mixed treatment comparisons. *BMC Med Res Methodol.* 2012;12(1):138.
 41. Schunemann H, Brozek J, Guyatt G, Oxman A. *GRADE Handbook* [Internet]. [cited 2023 Apr 6]. Available from: <https://training.cochrane.org/resource/grade-handbook>
 42. What is GRADE? | *BMJ Best Practice* [Internet]. [cited 2023 Apr 4]. Available from: <https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/>
 43. Chaimani A, Salanti G. Using network meta-analysis to evaluate the existence of small-study effects in a network of interventions. *Res Synth Methods.* 2012;3(2):161–76.

Enhanced Recovery After Surgery (ERAS)

33

Kathryn J. Uhlman, Alexandra C. Hatchell,
Achilles Thoma, and Claire F. Temple-Oberle

Abbreviations

ERAS	Enhanced Recovery After Surgery
LOS	Length of Stay
GDG	Guideline Development Group
PICOT	Population, Intervention, Comparator, Outcome, and Time Horizon
RCT	Randomized Control Trial

Introduction

The concept of Enhanced Recovery After Surgery (ERAS®) has been used to standardize optimal perioperative care pathways [1]. ERAS guidelines encompass evidence-based, multidisciplinary, and multifaceted components within the preoperative, intraoperative, and postoperative time period, with the goal to optimize patient recovery [2]. Shortly after the concept of ERAS

was first introduced by Henrik Kehlet in 1997, the ERAS® Society for perioperative care was established [1, 3]. The society was formed with the aim to develop evidence-based perioperative care through research, education, audit, and implementation [4]. ERAS® Society consensus statements and guidelines in multiple disciplines have been published [5–10].

ERAS guidelines are proven to improve patient outcomes and satisfaction, while reducing complications, recovery time, length of hospital stay (LOS) and total cost [2]. This has led to an increased number of “ERAS” guideline publications and “ERAS” implementation studies [11]. However, there is confusion in regard to the methodological and reporting structure that should be followed in the development of an ERAS guideline [11]. Despite the term *ERAS* being widely used by many authors, many published ERAS guidelines are not endorsed by the

K. J. Uhlman
Division of Plastic, Reconstructive and Aesthetic
Surgery, Department of Surgery, University
of Toronto, Toronto, ON, Canada
e-mail: kate.uhlman@medportal.ca

A. C. Hatchell · C. F. Temple-Oberle (✉)
Department of Surgery, University of Calgary,
Calgary, AB, Canada

Department of Oncology, University of Calgary,
Calgary, AB, Canada
e-mail: alexandra.hatchell@medportal.ca;
claire.temple-oberle@ahs.ca

A. Thoma
Division of Plastic Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: athoma@mcmaster.ca

ERAS® Society. Moreover, many published ERAS studies do not have adequate methodology or reporting quality [12].

To assist authors in the development of ERAS protocols, the ERAS® Society has published ERAS guideline development recommendations [11]. Additionally, *The Reporting on ERAS Compliance, Outcomes, and Elements Research (RECOVER) Checklist* [13] has been published in a joint statement between ERAS® and ERAS® USA Societies, and is included in the EQUATOR Network [14], an organization aimed to improve the quality of research and publications. With increased interest in ERAS guidelines, it is critical that readers are equipped with tools and skills to appraise the quality of an ERAS guideline prior to its adoption to their practice [15]. The purpose of the appraisal recommendations provided in this chapter is to equip the reader with a user's guide for the interpretation and application of ERAS guidelines.

Clinical Scenario

You are a newly recruited bariatric surgeon in a community hospital and wish to implement an ERAS protocol for bariatric surgery at your institution. You experience resistance from some senior surgeons and anesthesiologists who claim the care they provide for bariatric patients is already exemplary and do not wish to add more bureaucracy to their clinical practice. The chief of surgery is keen to support you but has some reservations as to the quality of published ERAS guidelines in other surgical specialties. They challenge you to find a published bariatric surgery ERAS protocol and perform a critical appraisal, and if happy with the quality of the ERAS guideline, they will recommend its implementation to other stakeholders (i.e., department of anesthesia and nursing).

Literature Search

You review the users' guide article on how to perform a high-quality literature search [16] (see also Chap. 4) and use MEDLINE [17] to com-

plete the search strategy (Appendix). The search yields 115 articles.

You screen the articles and narrow the results to only those articles that include a bariatric surgery ERAS guideline ($n = 53$). Of those, most ($n = 44$) were ERAS implementation studies. Some ($n = 3$) ERAS guidelines were specific to other populations such as pediatrics or patients with specific comorbidities. Four bariatric surgery ERAS guidelines were found to be endorsed by national organizations (Canada [18], Czech Republic [19], Italy [20], and an international organization) [21]. The only ERAS protocol endorsed by the ERAS® Society is the article, *Guidelines for Perioperative Care in Bariatric Surgery: Enhanced Recovery After Surgery (ERAS) Society Recommendations: A 2021 Update* [21] (2021 ERAS Guidelines for Perioperative Care in Bariatric Surgery) by Erik Stenberg, Luiz Fernando dos Reis Falca, Mary O'Kane, Ronald Liem, Dimitri J. Pournaras, Paulina Salminen, Richard D. Urman, Anupama Wadhwa, Ulf O. Gustafsson, and Anders Thorell.

You proceed to appraise this ERAS protocol following the questions shown in Box 33.1.

Box 33.1: Guidelines to Appraise an Article That Claims to Have Developed an ERAS Guideline

Primary Guides

Primary Guide I: Did the authors define the scope and purpose of the ERAS® protocol?

1. Were the overall objective(s) and health question(s) clearly described?
2. Was the population to whom the guideline is meant to apply clearly described?

Primary Guide II: Did the authors abide by the following recommendations for the methodology and standardization of guideline development?

1. Did the authors include the ERAS® Society and form a guideline development group (GDG) with appropriate and relevant stakeholders?

(continued)

Box 33.1 (continued)

2. Was the literature search adequate?
3. Was the quality of evidence analyzed for each intervention of the ERAS guideline?
4. Have the authors elicited components of the ERAS guideline through a consensus method?

Primary Guide III: Does the ERAS guideline contain necessary components?

1. Does the final ERAS guideline include multimodal and interprofessional elements?

Secondary Guides

Secondary Guide I: Does the ERAS guideline contain necessary reporting elements?

1. Are the ERAS guideline recommendations specific and easily identifiable?
2. Are there any potential conflicts of interest (COI)?
3. Were the guidelines made available both in a peer-reviewed journal and on the ERAS Society website?

Secondary Guide II: Will this ERAS guideline help our organization?

1. Are there studies assessing the effect of the implementation of the ERAS guideline and can I adopt this ERAS guideline to my hospital?

instrument designed to assess the quality of clinical practice guidelines [22–24]. Using this instrument to appraise the scope and purpose of an ERAS protocol, the following questions should be addressed:

1. *Were the overall objective(s) and health question(s) clearly described?*

Within the opening paragraphs of the article, the overall objective of the ERAS guideline should be described, and the expected health benefits should be outlined [22]. The health questions should be listed and provide the following details of the ERAS protocol: health objective(s), intervention, time period, outcomes, and health-care setting [15, 22]. It should be considered whether these elements are well-written, clear, easy to find, and applicable to the reader [22].

The *2021 ERAS Guidelines for Perioperative Care in Bariatric Surgery* [21] article includes the overall objective and expected health benefit within the first two paragraphs. The authors clearly describe that ERAS guideline implementation can reduce morbidity rates, shorten functional recovery, and reduce length-of-stay (LOS) in bariatric surgery. The health questions are sufficiently addressed within the introduction.

2. *Was the population to whom the guideline is meant to apply clearly described?*

The explanation of target population(s) should include details of the: (1) age and sex; (2) clinical condition; (3) severity of condition; (4) comorbidities; and (5) populations not included [15, 22].

Within the *2021 ERAS Guidelines for Perioperative Care in Bariatric Surgery* [21], the population to whom the guideline is meant to apply was clearly outlined as patients undergoing bariatric surgery. Although the authors did not outline specific details of age or sex, severity, or comorbidities, it can be implied that all individuals undergoing bariatric surgery are eligible participants in the ERAS protocol.

Article Appraisal

Primary Guide I: Did the Authors Define the Scope and Purpose of the ERAS© Protocol?

To appraise the scope and purpose of an ERAS protocol, elements from the AGREE II Checklist [22] can be used. The AGREE II Checklist is an

Primary Guide II: Did the Authors Abide by the Following Recommendations for the Methodology and Standardization of Guideline Development?

1. *Did the authors include the ERAS® Society and form a guideline development group (GDG) with appropriate and relevant stakeholders?*

An important primary step in the development of an ERAS protocol is the formation of a guideline development group (GDG) [11]. The GDG should have international representation and should comprise one or two lead authors within a maximum 10-person team [11]. The team should be multidisciplinary, with representation including surgeons, anesthesiologists, and other experts (oncologists, medical specialists, nurses, nutritionists, professional body representatives, health economists, and patients) [11]. Consultation with other stakeholders throughout the process is another important element of ERAS protocol development [11].

The authors of the *2021 ERAS Guidelines for Perioperative Care in Bariatric Surgery* [21] consisted of a ten-member multidisciplinary group, with international representation from disciplines including general surgery, anesthesia, nutrition, and clinical sciences. However, it is unclear if patients were consulted in the development of the ERAS protocol. The involvement of patients or caregivers would offer insight into the perspective of the appropriateness and feasibility of guidelines for patients and their loved ones [25].

2. *Was the literature search adequate?*

Possible recommendations to include within an ERAS guideline are first identified by the GDG through a literature search [11]. Establishing a well-developed research question is the primary step in a literature search. The development of an appropriate research question can be accomplished through the utilization of the PICOT (Population, Intervention, Comparator, Outcome, and Time

Horizon) format [16]. After the research question is established, inclusion/exclusion criteria are defined and target outcomes are identified [11]. Included resources should be reviewed by the GDG; this process should be highlighted within the ERAS article with a PRISMA diagram [11, 26].

The literature search conducted for the *2021 ERAS Guidelines for Perioperative Care in Bariatric Surgery* [21] was built from the literature search that had been previously conducted for the original bariatric surgery ERAS guidelines published in 2016 [27]. Although specific details of the research question are not outlined, process for inclusion criteria and target outcomes are detailed in both the 2016 [27] version, and the 2021 update [21]. No PRISMA [26] diagram was used within either article, however, the search strategy was depicted in detail, including utilized databases, search terms, and secondary resources as references of relevant studies.

3. *Was the quality of evidence analyzed for each intervention of the ERAS guideline?*

The recommendations included within an ERAS guideline should be assessed for quality of evidence and strength of recommendations. The ERAS® Society recommendations [11] include the utilization of the *Grading of Recommendations, Assessment, Development and Evaluation* (GRADE) [28] approach to appraise the quality of evidence of each ERAS guideline item. The GRADE approach offers a system of evaluating evidence and incorporating consideration of patient and society values to arrive at recommendations. The GRADE approach determines the level of evidence for each ERAS recommendation as *high, moderate, low, or very low* [28]. Once the GRADE level of evidence has been completed, the strength of each recommendation is determined and labeled as *strong or weak* [11, 28]. The clinical effectiveness, cost-effectiveness, and impact to quality of life should also be analyzed and commented on by the GDG [11].

The *2021 ERAS Guidelines for Perioperative Care in Bariatric Surgery* [21] evaluated the quality of evidence for each recommendation

through a systematic method involving multiple authors. The GRADE system was used to appraise the level of evidence and strength of each recommendation. This process was clearly described in the methods section of the study. Clinical effectiveness was commented upon, but cost-effectiveness and impact on quality of life were not expanded on within the article.

4. *Have the authors elicited components of the ERAS guideline through a consensus method?*

The Delphi process can be used to resolve discrepancy of agreement on potential ERAS guideline items [11]. The Delphi process is a stepwise method designed to obtain the most reliable consensus within a group of experts [29]. Key features of the Delphi method include: anonymity, iteration, controlled feedback, and statistical aggregation [29]. The process can operate through the distribution of structured, anonymous questionnaires. Each item is represented through a statistical average in the final round [29].

The GDG of the *2021 ERAS Guidelines for Perioperative Care in Bariatric Surgery* [21] used the Delphi method to reach consensus when there was disagreement regarding the appraisal of the strength of a recommendation. The ERAS® Society recommendations on ERAS guideline development include the recommendation to GDGs to develop a large number of proposed ERAS elements that are then voted on for inclusion, with consensus achieved through the Delphi method [11]. The authors of the *2021 ERAS Guidelines for Perioperative Care in Bariatric Surgery* [21] likely did not require this step, as the ERAS protocol was informed by the 2016 version [27].

Primary Guide III: Does the ERAS Guideline Contain Necessary Components?

1. *Does the final ERAS guideline include multimodal and interprofessional elements?*

It is integral for ERAS guideline elements to be multimodal and interdisciplinary [1, 11, 13].

Multimodal care is the combination of multiple interventions, rehabilitation, care programs, and other therapeutic modalities [30]. Multimodal care includes multiple interventions across the care pathway (preoperative, intraoperative, and postoperative) [30]. Additionally, one management component can encompass multimodal care elements. For example, a multimodal analgesia strategy is a key component to many ERAS protocols [31].

The *2021 ERAS Guidelines for Perioperative Care in Bariatric Surgery* [21] are multimodal and include recommendations across the entire care pathway (preadmission, preoperative, intraoperative, and postoperative). Examples of preadmission elements include information, education and counseling, smoking and alcohol cessation, preoperative weight loss, etc. Within the perioperative time period, examples include preoperative fasting, intraoperative standardized anesthetic protocol, and early postoperative nutritional care. Additionally, there are multimodal elements within interventions, such as multimodal postoperative nausea and vomiting (PONV) prophylaxis and multimodal analgesia.

Secondary Guides

Secondary Guide I: Does the ERAS Guideline Contain Necessary Reporting Elements?

1. *Are the ERAS guideline recommendations specific and easily identifiable?*

Each recommendation within an ERAS guideline should be concrete, precise, and evidence-based [22–24]. The fourth domain of the AGREE II checklist [22] can be used for the appraisal of the specificity and identifiability of recommendations within an ERAS guideline. Each recommendation within an ERAS guideline should include: (1) a statement of recommended action; (2) identification of the intent or purpose of the recommended action; (3) qualifying statements, if relevant; and (4) identification of the relevant population [22].

Each recommendation within the 2021 *ERAS Guidelines for Perioperative Care in Bariatric Surgery* [21] is clear and precise. Within the *Results: Evidence base and recommendations* section of the article, there are headings with the statement of recommendation, followed by details of the intent and purpose of each recommended guideline item. For example, under the heading of the preadmission item, *information, education and counseling*, the intent of the guideline item is explained as, “to ensure realistic expectations, and to reduce anxiety, wound complications, postoperative pain and LOS.” Qualifying statements are used when appropriate, and high-quality studies are cited throughout to support statements made. For example, under the heading of the preadmission item, *preoperative weight loss*, the positive results of several observational studies were referenced, with the qualifying statement that the effect of preoperative weight loss had not been addressed in any randomized control trial (RCT) or large retrospective study. The relevant population is outlined in the *indications and contraindications for surgery* section of the article.

2. Are there any potential conflicts of interest (COI)?

The 6th domain of the AGREE II checklist [22] can be used to assess editorial independence. An ERAS guideline should report the following: (1) disclosure and influence of funding body (if applicable); and (2) competing interests for each GDG member [22].

The 2021 *ERAS Guidelines for Perioperative Care in Bariatric Surgery* [21] article does include a description of the funding bodies that supported the study. These included Orebro University, and grants from Region Orebro County (Grant No. OLL-939106), the Bengt Ihre Foundation, and the Erling Persson Family Foundation (Grant No. 140604). Conflicts of interest are disclosed for each author at the end of the article. The funding bodies and conflicts of interest likely did not influence the results or the final recommendations of the ERAS guideline.

3. Were the guidelines made available both in a peer-reviewed journal and on the ERAS® Society website?

A requirement of an ERAS® Society Guideline is for the protocol to be endorsed by the ERAS® Society leadership [11]. Recommendations from the ERAS® Society on the development of an ERAS guideline include the recommendation for authors to engage the ERAS® Society early-on in the guideline development process [11]. Alternatively, in the final stages of the ERAS guideline development process, prior to submission to a journal, the authors can negotiate endorsement by the ERAS® Society so that the guidelines are made publicly available on the ERAS® Society website. An ERAS® Society endorsement reassures readers that the ERAS guideline is of high methodological and reporting quality. The ERAS® Society has a process of reviewing guidelines every 2–3 years, therefore ensuring up-to-date ERAS guideline resources [4].

The 2021 *ERAS Guidelines for Perioperative Care in Bariatric Surgery* [21] are endorsed by the ERAS® Society and are available on the ERAS® Society website [4]. The article was published in the *World Journal of Surgery* [21]. In the 2016 version, the article includes a description that the ERAS guideline was an initiative by the ERAS® Society, and was endorsed by the International Association for Surgical Metabolism and Nutrition [32].

Secondary Guide II: Will this ERAS Guideline Help Our Organization?

1. Are there studies assessing the effect of the implementation of the ERAS guideline and can I adopt this ERAS guideline to my hospital?

When deciding whether to adopt an ERAS guideline at an institution, it is important to review the literature for studies evaluating impact of the guideline. To assist clinician-scientists in the reporting of outcomes of an ERAS in practice, the ERAS® Society published *The Reporting*

on *ERAS Compliance, Outcomes, and Elements Research (RECOVER) Checklist* [13]. Readers can refer to the RECOVER checklist when evaluating the reporting quality of an ERAS implementation study.

The *2021 ERAS Guidelines for Perioperative Care in Bariatric Surgery* [21] article is not an ERAS implementation study and therefore, the *RECOVER checklist* does not apply. A review of the literature reveals many bariatric surgery ERAS implementation studies. A systematic review [33] included five randomized controlled trials (RCTs) and 12 observational studies that compared the bariatric surgery ERAS protocol with standard care. The study found the ERAS protocol led to decreased length of the hospital stay ($p < 0.01$), and decreased incidence of PONV ($p < 0.01$) [33]. The *RECOVER checklist* can be used to assess the reporting quality of each of the included studies within the systematic review.

Resolution of Clinical Scenario

You present the *2021 ERAS Guidelines for Perioperative Care in Bariatric Surgery* to the chief of surgery and explain that overall, it is of excellent methodological quality. Key strengths include the methodology of guideline development, including a multidisciplinary and international GDG, with ERAS® Society involvement. The standardization of guideline development was closely followed, with the incorporation of quality of evidence appraisal and consensus method. The ERAS guideline included multi-modal and interprofessional elements. Necessary reporting elements are also included. Implementation studies published in the literature have shown the implementation of the ERAS protocol to have positive patient outcomes. The chief of surgery is pleased with the quality of this ERAS protocol and decides to follow your recommendation. They will recommend the adoption of this ERAS protocol for bariatric patients to the chiefs of anesthesia and nursing at your institution.

Appendix: Search Strategy (Date of Search: February 27, 2024)

1	Exp bariatric surgery/
2	Exp Roux-en-Y*/ or exp gastric bypass*/ or exp sleeve gastrectomy*/ or exp biliopancreatic diversion with duodenal switch*/ or BPD*/ or *DS*/ or exp single-anastomosis duodeno-ileal bypass with sleeve gastrectomy*/ or SADI-S*/ or exp adjustable gastric band*/ or AGB*
3	1 OR 2
4	(Enhanced Recovery After Surgery* or ERAS* or fast track surgery* or fast recovery*).ti,ab,kw.
5	3 AND 4
6	Limit to English
7	Limit to 2018 to February 27, 2024

References

1. Ljungqvist O, Young-Fadok T, Demartines N. The history of enhanced recovery after surgery and the ERAS society. *J Laparoendosc Adv Surg Tech A*. 2017;27:860–2.

2. Lau CSM, Chamberlain RS. Enhanced recovery after surgery programs improve patient outcomes and recovery: a meta-analysis. *World J Surg*. 2017;41:899–913.

3. Golder HJ, Papalois V. Enhanced recovery after surgery: history, key advancements and developments in transplant surgery. *J Clin Med*. 2021;10:1634.

4. Home – ERAS® Society. <https://erassociety.org/>. Accessed 26 Feb 2024.

5. Temple-Oberle C, Shea-Budgell MA, Tan M, Semple JL, Schrag C, Barreto M, Blondeel P, Hamming J, Dayan J, Ljungqvist O. Consensus review of optimal perioperative care in breast reconstruction: enhanced recovery after surgery (ERAS) society recommendations. *Plast Reconstr Surg*. 2017;139:1056e–71e.

6. Dort JC, Farwell DG, Findlay M, Huber GF, Kerr P, Shea-Budgeli MA, Simon C, Uppington J, Zygun D, Ljungqvist O, Harris J. Optimal perioperative Care in Major Head and Neck Cancer Surgery with Free Flap Reconstruction: a consensus review and recommendations from the enhanced recovery after surgery society. *JAMA Otolaryngol Head Neck Surg*. 2017;143:292–303.

7. Nelson G, Fotopoulou C, Taylor J, Glaser G, Bakkum-Gamez J, Meyer LA, Stone R, Mena G, Elias KM, Altman AD, Bisch SP, Ramirez PT, Dowdy SC. Enhanced recovery after surgery (ERAS®) society guidelines for gynecologic oncology: addressing implementation challenges-2023 update. *Gynecol Oncol*. 2023;173:58–67. NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

8. Batchelor TJP, Rasburn NJ, Abdelnour-Berchtold E, Brunellid A, Cerfolioe RJ, Gonzalezc M, Ljungqvistf

- O, Peterseng RH, Popescu WM, Slinger PD, Naidu B. Guidelines for enhanced recovery after lung surgery: recommendations of the Enhanced Recovery After Surgery (ERAS®) Society and the European Society of Thoracic Surgeons (ESTS). *Eur J Cardiothorac Surg.* 2019;55:91–115.
9. Cerantola Y, Valerio M, Persson B, Jichlinski P, Ljungqvist O, Hubner M, Kassouf M, Muller S, Baldini G, Carli F, Naesheim T, Ytrebo L, Revhaug A, Lassen K, Knutsen T, Aarsether E, Wiklund P, Patel HRH. Guidelines for perioperative care after radical cystectomy for bladder cancer: enhanced recovery after surgery (ERAS®) society recommendations. *Clin Nutr.* 2013;32:879–87.
10. McGinagle KL, Spangler EL, Ayyash K, Vavra AK, Arya S, Settembrini AM, Thomas MM, Dell KE, Swiderski IJ, Davies MG, Setacci C, Urman RD, Howell SJ, Garg J, Ljungqvist O, De Boer HD. A framework for perioperative care for lower extremity vascular bypasses: a Consensus Statement by the Enhanced Recovery after Surgery (ERAS®) Society and Society for Vascular Surgery. *J Vasc Surg.* 2023; <https://doi.org/10.1016/j.jvs.2023.01.018>.
11. Brindle M, Nelson G, Lobo DN, Ljungqvist O, Gustafsson UO. Recommendations from the ERAS® Society for standards for the development of enhanced recovery after surgery guidelines. *BJS Open.* 2020;4:157–63.
12. Uhlman K, Behroozian T, Lewandowski N, Yuan M, Thoma A. The quality of plastic surgery Enhanced Recovery After Surgery (ERAS) studies: a systematic review. *J Plast Reconstr Aesthet Surg.* 2024; <https://doi.org/10.1016/j.bjps.2024.11.0>.
13. Elias KM, Stone AB, McGinagle K, Tankou JI, Scott MJ, Fawcett WJ, Demartines N, Lobo DN, Ljungqvist O, Urman RD. The reporting on ERAS compliance, outcomes, and elements research (RECOVER) checklist: a joint statement by the ERAS® and ERAS® USA societies. *World J Surg.* 2019;43:1–8.
14. EQUATOR Network | Enhancing the QUALity and Transparency Of Health Research. <https://www.equator-network.org/>. Accessed 26 Feb 2024.
15. Uhlman K, Sardiwalla Y, Chin B, Murphy J, Finley C, Thoma A. Users' guide to the surgical literature: how to appraise ERAS guidelines. *J Cardiothorac Vasc Anesth.* 2022;36:2857–63.
16. Waltho D, Kaur MN, Haynes RB, Farrokhyar F, Thoma A. Users' guide to the surgical literature: how to perform a high-quality literature search. *Can J Surg.* 2015;58:349.
17. Ovid – Ovid MEDLINE® | Wolters Kluwer. <https://www.wolterskluwer.com/en/solutions/ovid/ovid-medline-901>. Accessed 26 Feb 2024.
18. Dang JT, Szeto VG, Elnahas A, Ellsmere J, Okrainec A, Neville A, Malik S, Yorke E, Hong D, Biertho L, Jackson T, Karmali S. Canadian consensus statement: enhanced recovery after surgery in bariatric surgery. *Surg Endosc.* 2020;34:1366–75.
19. Satinský I, Hrubý M, Šrámková P, Patka J, Čierný M, Babiak P, Šimonik I, Schwarz P, Haluzík M. Clinical practice guidelines for perioperative care in bariatric surgery 2023: adapted ERAS (enhanced recovery after surgery) guidelines with consensual voting of the working group of the Joint Bariatric and Metabolic Surgery Section of the Czech Surgery.... *Rozhl Chir.* 2024;102:283–97.
20. Marinari G, Foletto M, Nagliati C, Navarra G, Borrelli V, Bruni V, Fantola G, Moroni R, Tritapepe L, Monzani R, Sanna D, Carron M, Cataldo R. Enhanced recovery after bariatric surgery: an Italian consensus statement. *Surg Endosc.* 2022;36:7171–86.
21. Stenberg E, dos Reis Falcão LF, O'Kane M, Liem R, Pournaras DJ, Salminen P, Urman RD, Wadhwa A, Gustafsson UO, Thorell A. Guidelines for perioperative Care in Bariatric Surgery: Enhanced Recovery After Surgery (ERAS) Society Recommendations: a 2021 update. *World J Surg.* 2022;46:729.
22. AGREE Next Steps Consortium. The AGREE II Instrument [Electronic version] (2017). Retrieved December 18, 2024, from <http://www.agreetrust.org>.
23. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, Fervers B, Graham ID, Hanna SE, Makarski J. Development of the AGREE II, part 1: performance, usefulness and areas for improvement. *CMAJ.* 2010;182:1045–52.
24. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, Fervers B, Graham ID, Grimshaw J, Hanna SE, Littlejohns P, Makarski J, Zitzelsberger L. AGREE II: advancing guideline development, reporting and evaluation in health care. *CMAJ. Can Med Assoc J.* 2010;182:E839.
25. Armstrong MJ, Mullins CD, Gronseth GS, Gagliardi AR. Impact of patient involvement on clinical practice guideline development: a parallel group study. *Implement Sci.* 2018;13:1–13.
26. PRISMA. <http://prisma-statement.org/prismastatement/flowdiagram.aspx?AspxAutoDetectCookieSupport=1>. Accessed 26 Feb 2024.
27. Thorell A, MacCormick AD, Awad S, Reynolds N, Roulin D, Demartines N, Vignaud M, Alvarez A, Singh PM, Lobo DN. Guidelines for perioperative care in bariatric surgery: Enhanced Recovery After Surgery (ERAS) Society Recommendations. *World J Surg.* 2016;40(9):2065–83.
28. Guyatt GH, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schünemann HJ. GRADE guidelines: 1. Introduction – GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol.* 2011;64:383–94.
29. Okoli C, Pawlowski SD. The Delphi method as a research tool: an example, design considerations and applications. *Inf Manag.* 2004;42:15–29.
30. National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Health Sciences Policy. Forum on Neuroscience and Nervous System Disorders. Developing multimodal therapies for brain disorders: proceedings of a workshop. Washington, DC: National Academies Press (US); 2016. PMID: 27875040.

31. Gedda C, Nygren J, Garpenbeck A, Hoffström L, Thorell A, Soop M. Multimodal analgesia bundle and postoperative opioid use among patients undergoing colorectal surgery. *JAMA Netw Open*. 2023;6:e2332408.
32. AISMEN International Association for Surgical Metabolism and Nutrition. <https://www.iasmen.org/>. Accessed 26 Feb 2024.
33. Zhou J, Du R, Wang L, Wang F, Li D, Tong G, Wang W, Ding X, Wang D. The application of Enhanced Recovery After Surgery (ERAS) for patients undergoing bariatric surgery: a systematic review and meta-analysis. *Obes Surg*. 2021; <https://doi.org/10.1007/s11695-020-05209-5>/Published.

Appraising Core Outcome Sets (COS) in Surgical Research

34

George Elzawy, Achilles Thoma,
Yaeesh Sardiwalla, Jessica Murphy,
Marta Karpinski, and Richard P. Whitlock

Abbreviations

COS	Core outcome sets
COMET	Core outcome measures in effectiveness trials
RCT	Randomized controlled trial
TAVR	Transcatheter aortic valve replacement
SAVR	Surgical aortic valve replacement
COS-STAD	The Core Outcome Set-STANDards for Development
CABG	Coronary artery bypass grafts
OMERACT	Outcome measures in rheumatoid arthritis clinical trials initiative
MACE	Major adverse cardiac events

Introduction

Knowledge translation is predicated on the accuracy of methodology and reporting in clinical research [1]. Systematic reviews and meta-analyses facilitate knowledge translation through synthesizing existing evidence and identifying gaps in knowledge to guide future research and clinical practice [2–5]. By pooling data from multiple small studies, one increases the statistical power to detect differences between interventions [4, 5]. As such, systematic reviews and meta-analyses can provide the highest level of evidence for informing clinical practice (see Chap. 16). A prerequisite for pooling data is con-

G. Elzawy · M. Karpinski
Temerty Faculty of Medicine, University of Toronto,
Toronto, ON, Canada
e-mail: george.elzawy@mail.utoronto.ca;
marta.karpinski@mail.utoronto.ca

A. Thoma (✉)
Division of Plastic Surgery, Department of Surgery
& Department of Health Research Methods,
Evidence and Impact, McMaster University,
Hamilton, ON, Canada
e-mail: athoma@mcmaster.ca

Y. Sardiwalla
Division of Plastic Surgery, Department of Surgery,
McMaster University, Hamilton, ON, Canada
e-mail: yaeesh.sardiwalla@medportal.ca

J. Murphy
Division of Physical Medicine & Rehabilitation,
Department of Medicine, McMaster University,
Hamilton, ON, Canada
e-mail: murphj11@mcmaster.ca

R. P. Whitlock
Department of Surgery, Division of Cardiac Surgery,
McMaster University, Hamilton, ON, Canada
Department of Health Research Methods, Evidence,
and Impact, McMaster University,
Hamilton, ON, Canada

Population Health Research Institute,
Hamilton, ON, Canada
e-mail: richard.whitlock@phri.ca

sistent outcome measurement and reporting in primary studies. Unfortunately, individual studies investigating the same intervention often measure vastly different outcomes [6, 7]. In other instances, even when the same outcomes are assessed between studies, there is often inconsistency in the way they are defined and measured, making it difficult, or impossible to pool data. Inconsistent outcome selection, measurement, and reporting contributes to research inefficiency and raises the question of which outcomes are most important to measure in clinical research.

The Core Outcome Measures in Effectiveness Trials (COMET) Initiative advocates for the standardization of outcomes in research through the implementation of Core Outcome Sets (COS) [8]. The development and use of COS may help address the problem of inconsistent outcome selection and reporting. The development of a COS involves gathering outcomes reported in the literature, organizing these outcomes into domains, and then identifying the most critical outcomes through a consensus method with key stakeholders (e.g., patients, surgeons, and caregivers) [9]. The finalized COS is a list of the most critical outcomes that should be measured and reported in all clinical trials investigating a particular health condition [9]. Therefore, the application of a COS in clinical research promotes the measurement and reporting of the most salient outcomes, ensuring that results will be amenable to pooling in future evidence syntheses. In effect, the use of a COS helps increase the likelihood that the results of clinical research studies will be meaningfully used to inform clinical decision-making.

As many surgical specialties are introducing COS for various conditions, the purpose of this chapter is to help surgeons and researchers appraise studies that claim to have developed a COS. Through this appraisal, surgeon-investigators should be able to select a valid COS to inform study outcomes in their own research.

Clinical Scenario

You are a newly recruited cardiac surgeon at an academic center. You are interested in undertaking a large Randomized Controlled Trial (RCT) to compare transcatheter aortic valve replacement (TAVR) and surgical aortic valve replacement (SAVR). However, in your review of the literature, you find there are many reported outcomes in cardiac surgery. This problem has been the subject of heated discussions at your Division's monthly journal clubs. The discussions pertained to the multiplicity of chosen outcomes in similar studies and consequently, the inability to include these articles in systematic reviews or meta-analyses. As you intend to apply for a large competitive peer-reviewed grant, you want to ensure that your RCT outcomes are clinically important and defensible. You are advised by a colleague that a COS may be helpful in determining the critical outcomes to measure in your planned RCT.

Literature Search

After referring to a previous Users' Guide article on performing a literature search [10], or the chapter in this book (see Chap. 4) you develop the following search strategy based on your research question: *adult AND cardiac surgery AND core outcome set OR COS*. You input your search terms into the Cochrane library database and restrict your search to studies published in the last 10 years (2014–2024). The search yields ten results (see [Appendix](#)). Of the ten studies, the only study that claims to have developed a COS is “A core outcome set for adult cardiac surgery trials: A consensus study,” by Benstoem and colleagues [11]. You proceed with appraising this COS using questions inspired by the COMET handbook (Box 34.1) [9]. Readers are encouraged to visit this reference for greater detail.

Box 34.1: Guidelines to Appraise an Article That Claims to Have Developed a Core Outcome Set

A. Is this COS valid?

Primary guides:

- (i) Did the authors define the scope of the COS?
- (ii) Did the authors explain the need for a COS in the specific condition?
- (iii) Have the authors developed and adhered to a protocol to achieve consensus?
 - (a) Did the authors utilize multiple sources to identify relevant outcomes?
 - (b) Were relevant stakeholders involved?
 - (c) Have the authors elicited important outcomes through a consensus method?

Secondary guides:

- (i) Did the authors register the COS with the COMET database?
- (ii) Are there any potential conflicts of interest or biases?

B. What was the final agreed upon COS?

Were each of the core outcomes defined to permit reproducibility?

C. Can I apply this COS to my research?

Will this COS help me develop my RCT?

research or practice setting to which the COS is to be applied. Benstoem and colleagues [11] explicitly specified the population (adults >18 years of age), the research setting (the COS is to be applied only to effectiveness trials), and the intervention (invasive on- or off-pump cardiac surgery); the health condition (ischemic heart disease) is implied in the article's introduction. An important component of your TAVR versus SAVR study is the intervention; you are comparing an invasive surgery to a minimally invasive procedure. Cardiac research presently reflects this trend as more RCTs are being published to study the effectiveness of minimally invasive procedures compared to invasive surgeries [12]. For example, TAVRs are suitable alternatives to SAVRs¹³ and stent angioplasty is an option for revascularization as opposed to coronary artery bypass grafts (CABG) [14]. The COS developed by Benstoem and colleagues [11] is to be used in studies that compare invasive cardiac procedures (e.g., comparing on- and off-pump CABG), and so you wonder if this COS is appropriate for RCTs comparing minimally invasive and invasive procedures. From the scope of the COS [11], you make the decision that, moving forward in your appraisal, if valid, this COS would be beneficial to one arm of your effectiveness trial (the invasive arm, SAVR, outcomes). However, you will need to support the other arm (TAVR) with TAVR-relevant outcomes; that is, outcomes relevant to minimally invasive cardiac procedures.

- (ii) *Did the authors explain the need for a COS in the specific condition?*

The first step of identifying a need for a COS is to prove that there is heterogeneity in outcome reporting for a particular condition. The COMET handbook [9] suggests reviewing previous trials or systematic reviews to identify the need for a COS. The need for a COS is justified when there is a large number of outcomes reported within the literature. Benstoem and colleagues [11] address this in their COS protocol, a separate publication [12], analyzing 15 systematic reviews. This study [12], found that most of the systematic reviews

A. Is This Core Outcome Set Valid?

Primary Guides:

- (i) *Did the authors define the scope of the COS?*

One of the most important components of a scholarly article is clear communication of who can benefit from the information. According to the Core Outcome Set-STAndards for Development (COS-STAD) recommendations, there are four specifications to properly define the scope of a COS: (1) the health condition; (2) the population; (3) the intervention; and (4) the

were unable to pool quantitative RCT data into a meta-analysis because of absent data or heterogeneous outcome measures.

The second step of identifying the need for a COS is to ensure that there is either no established COS or that the existing COS is unsatisfactory or inapplicable to your study. The COMET database provides a comprehensive list of previously published COS and is updated annually [9]. This database allows researchers to search for established or ongoing COS studies for any given condition. Using the COMET database, you confirm that the COS by Benstoem and colleagues [11] is the first and only COS in adult cardiac surgery.

(iii) *Have the authors reported and adhered to a protocol to achieve consensus?*

It is recommended that a protocol is published prior to producing the COS study [9]. Subsequent adherence to this protocol is critical to executing a COS study that aligns with current recommendations and mitigating biases that arise from significant changes to the protocol. The following three subsidiary questions should be answered when appraising the consensus methods used in the development of a COS.

(a) *Did the authors utilize multiple sources to identify relevant outcomes?*

As discussed above, when developing a COS, one must first identify outcomes found in existing literature. These extracted outcomes, in combination with new outcomes identified by stakeholders, inform the development of a COS. The COMET handbook⁹ recommends four sources that should be considered when collecting existing outcomes: systematic reviews, reviews of qualitative studies, national audit data sets, and interviews or focus groups with relevant stakeholders. When all four data sources are utilized, there is a more representative consensus in the COS development process. The article you are appraising [11] relied on systematic reviews to extract potential outcomes. Without the use of other sources, there is a higher risk of missing

other outcomes not yet measured in cardiac surgery research that may be important to stakeholders.

(b) *Were relevant stakeholders involved?*

Stakeholders are involved in the “consensus process” stage of COS development (see below). The stakeholders involved in COS development should represent those groups of individuals who will ultimately use or benefit from the results of clinical research in the given area. Benstoem and colleagues [11] included cardiothoracic surgeons, patients, anesthesiologists, nursing staff, researchers with expertise in the condition, cardiologists, perfusionists, and physicians in their consensus process. Perhaps more important than quantitative representation is proportional representation; even if patients are present in the consensus procedure, disproportional representation in favor of healthcare providers can negate patients’ perspectives. Benstoem and colleagues [11] had a 1:1 ratio of patients to cardiothoracic surgeons, thereby ensuring adequate representation. Additionally, rather than combining all perspectives, which risks missing discordance between patients and other stakeholders, the authors [11] astutely analyzed patients’ rankings separately from those of the healthcare providers. Thus, Benstoem and colleagues [11] were able to involve key stakeholders and emphasize the importance of patients’ priorities through separate analyses.

(c) *Have the authors elicited important outcomes through a consensus method?*

Common methods used to rank outcome domains in the production of a COS include: the Delphi survey, the nominal group technique, the consensus development conference, the semi-structured group discussion, or a combination of these [9]. Although there is sparse empirical evidence that different methods lead to different outcomes, the Delphi technique is frequently used [9]. In the Delphi technique, stakeholders complete anonymous sequential rounds of questionnaires to rank outcomes by perceived importance.

A threshold is typically defined in the COS development protocol to determine the minimum average rank of an outcome that permits it to proceed to the next “round,” while lower-ranked outcomes are excluded [9]. Benstoem and colleagues [11] followed the proper Delphi technique outlined in the COMET handbook [9] and clearly described their exclusion threshold (60% of stakeholders must rank the outcome as unimportant and 20% of stakeholders must rank the outcome as important to prevent the outcome from moving to the next round).

Secondary Guides

- (i) *Did the authors register the COS with COMET?*

Benstoem and colleagues [11] report registering their COS with COMET, this is confirmed by navigating the COMET database (www.comet-initiative.org/Studies). Therefore, you are satisfied that this precondition has been met.

- (ii) *Are there any potential conflicts of interest or biases?*

Similar to any study, in COS studies, conflicts of interests or biases should be disclosed by the authors. Studies that develop COS are especially susceptible to selection and attrition bias due to the extensive involvement required by stakeholders. Personalized correspondence, reducing the time commitment, and provision of clear instructions may help reduce these biases [15]. Benstoem and colleagues [11] report no competing interests and provide information regarding their funding from an internal program within their affiliated Faculty of Medicine. The authors also report their attempt at minimizing selection bias by inviting participants globally, although participation rates were still low [11].

B. What Was the Final Agreed Upon COS?

The authors stated in their protocol [12] that their COS, once formed, would involve one outcome for each of the four “core areas” recommended by the Outcome Measures in Rheumatoid

Arthritis Clinical Trials (OMERACT) Initiative: death, life impact, pathophysiological manifestations, and resource use [16]. Thus, the core outcomes were identified as *mortality*, *quality of life*, *cerebrovascular complications*, and *hospitalization*, respectively. Including these “core areas” represent stakeholder views; however, limiting one outcome per “core area” could have eliminated equally important outcomes. As an example, cerebrovascular complications were chosen as the core outcome to represent pathophysiological manifestations. However, measure of myocardial infarction was excluded despite an identical percentage of stakeholders voting “yes” in the Delphi surveys [11] and its importance in several cardiac interventions. In fact, combining cerebrovascular and cardiac outcomes is popular in cardiac surgery RCTs and is often represented as major adverse cardiac events (MACE) [17]. As such, further iterations of this cardiac COS can present outcomes as they are organized in the context of cardiac surgery research to maximize comprehensiveness.

C. Can I Apply This COS to My Research?

Your RCT aims to compare SAVR to TAVR, therefore, while the appraised COS [11] is useful for outcomes relevant to the SAVR arm, there are other important outcomes that need to be captured for TAVR. For example, when comparing TAVR and SAVR, limitations of TAVR include higher reintervention and rehospitalization rates (not included in this COS) [18].

Clinical Scenario Resolution

In summary, you deem the COS by Benstoem and colleagues [11] appropriate and sufficient for reporting the SAVR-specific outcomes you intend to measure in your TAVR versus SAVR RCT. In your grant proposal, you state your outcomes as mortality, quality of life, and hospitalization; you defend them using the COS by Benstoem and colleagues [11]. Rather than *cerebrovascular complications*, you include MACE as an outcome providing the justification stated above. To account for outcomes important to the comparison of TAVR and SAVR, you include outcomes

such as *rehospitalization* secondary to the procedure, heart failure, or valve dysfunction. Thus, your grant proposal includes baseline outcomes from the cardiac surgery COS [11] and TAVR/SAVR-comparative outcomes.

Through this appraisal, we can conclude that Benstoem and colleagues [11] followed most of the necessary steps to produce a valid COS. We have also highlighted the importance of defining core outcomes, such that, each surgeon-investigator designing their effectiveness trial appropriately understands and integrates these outcomes in their research. By now, surgeons should feel comfortable selecting an appropriate COS to support their research and make the necessary modifications to maximize the utility of a COS in their effectiveness trials.

Appendix

Search Results

1. Kotani Y, Kataoka Y, Izawa J, Fujioka S, Yoshida T, Kumasawa J, Kwong JS. High versus low blood pressure targets for cardiac surgery while on cardiopulmonary bypass. *Cochrane Database Syst Rev.* 2022;11 (11):CD013494. <https://doi.org/10.1002/14651858>.
2. Gibbison B, Villalobos Lizardi JC, Avilés Martínez KI, Fudulu DP, Medina Andrade MA, Pérez-Gaxiola G, Schadenberg AW, Stoica SC, Lightman SL, Angelini GD, Reeves BC. Prophylactic corticosteroids for paediatric heart surgery with cardiopulmonary bypass. *Cochrane Database Syst Rev.* 2020;10 (10):CD013101. <https://doi.org/10.1002/14651858.CD013101.pub2>.
3. Fabes J, Brunskill SJ, Curry N, Doree C, Stanworth SJ. Pro-coagulant haemostatic factors for the prevention and treatment of bleeding in people without haemophilia. *Cochrane Database Syst Rev.* 2018;12 (12):CD010649. <https://doi.org/10.1002/14651858.CD010649.pub2>.
4. Doiron KA, Hoffmann TC, Beller EM. Early intervention (mobilization or active exercise) for critically ill adults in the intensive care unit. *Cochrane Database Syst Rev.* 2018 Mar;3 (3):CD010754. <https://doi.org/10.1002/14651858.CD010754.pub2>.
5. Benstoem C, Stoppe C, Liakopoulos OJ, Ney J, Hasenclever D, Meybohm P, Goetzenich A. Remote ischaemic preconditioning for coronary artery bypass grafting (with or without valve surgery). *Cochrane Database Syst Rev.* 2017;5 (5):CD011719. <https://doi.org/10.1002/14651858.CD011719.pub3>.
6. Kidd EA, Stewart F, Kassis NC, Hom E, Omar MI. Urethral (indwelling or intermittent) or suprapubic routes for short-term catheterisation in hospitalised adults. *Cochrane Database Syst Rev.* 2015;2015 (12):CD004203. <https://doi.org/10.1002/14651858.CD004203.pub3>.
7. Lewis SR, Pritchard MW, Estcourt LJ, Stanworth SJ, Griffin XL. Interventions for reducing red blood cell transfusion in adults undergoing hip fracture surgery: An overview of systematic reviews. *Cochrane Database Syst Rev.* 2023;6 (6):CD013737. <https://doi.org/10.1002/14651858.CD013737.pub2>.
8. Burton JK, Craig LE, Yong SQ, Siddiqi N, Teale EA, Woodhouse R, Barugh AJ, Shepherd AM, Brunton A, Freeman SC, Sutton AJ, Quinn TJ. Non-pharmacological interventions for preventing delirium in hospitalised non-ICU patients. *Cochrane Database Syst Rev.* 2021;7 (7):CD013307. <https://doi.org/10.1002/14651858.CD013307.pub2>.
9. Liu Z, Dumville JC, Norman G, Westby MJ, Blazeby J, McFarlane E, Welton NJ, O'Connor L, Cawthorne J, George RP, Crosbie EJ, Rithalia AD, Cheng HY. Intraoperative interventions for preventing surgical site infection: an overview of Cochrane Reviews. *Cochrane Database Syst Rev.* 2018;2 (2):CD012653. <https://doi.org/10.1002/14651858.CD012653.pub2>.
10. Hegewald J, Wegewitz UE, Euler U, van Dijk JL, Adams J, Fishta A, Heinrich P, Seidler A. Interventions to support return to work for people with coronary heart disease. *Cochrane*

Database Syst Rev. 2019;3 (3):CD010748.
<https://doi.org/10.1002/14651858.CD010748.pub2>.

References

1. Chan AW, Song F, Vickers A, Jefferson T, Dickersin K, Gøtzsche PC, Krumholz HM, Ghersi D, van der Worp HB. Increasing value and reducing waste: addressing inaccessible research. *Lancet*. 2001;383:257–66. [https://doi.org/10.1016/S0140-6736\(13\)62296-5](https://doi.org/10.1016/S0140-6736(13)62296-5).
2. Haines T, McKnight L, Duku EK, Perry L, Thoma A. The role of systematic reviews in clinical research and practice. *Clin Plast Surg*. 2008;35:207–14. <https://doi.org/10.1016/j.cps.2007.10.003>.
3. Gopalakrishnan S, Ganeshkumar P. Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare. *J Family Med Prim Care*. 2013;2:9–14. <https://doi.org/10.4103/2249-4863.109934>.
4. Bangdiwala S, Bhargava A, O'Connor DP, Robinson TN, Michie S, Murray DM, Stevens J, Belle SH, Templin TN, Pratt CA. Statistical methodologies to pool across multiple intervention studies. *Transl Behav Med*. 2016;6:228–35. <https://doi.org/10.1007/s13142-016-0386-8>.
5. Bhandari M, Devereaux P, Montori V, Cinà C, Tandan V, Guyatt GH, Evidence-Based Surgery Working Group. Users' guide to the surgical literature: how to use a systematic literature review and meta-analysis. *Can J Surg*. 2004;47:60–7.
6. Copeland A, Gallo L, Weber C, Moltaji S, Gallo M, Murphy J, Axelrod D, Thoma A. Reporting outcomes and outcome measures in thumb carpometacarpal joint osteoarthritis: a systematic review. *J Hand Surg Am*. 2021;46:65.e1–65.e11. <https://doi.org/10.1016/j.jhsa.2020.05.024.42>.
7. Karpinski M, Moltaji S, Baxter C, Murphy J, Petropoulos JA, Thoma A. A systematic review identifying outcomes and outcome measures in Dupuytren's disease research. *J Hand Surg Eur*. 2020;45:513–20.
8. Prinsen CA, Vohra S, Rose MR, King-Jones S, Ishaque S, Bhaloo Z, Adams D, Terwee CB. Core Outcome Measures in Effectiveness Trials (COMET) initiative: Protocol for an international Delphi study to achieve consensus on how to select outcome measurement instruments for outcomes included in a "core outcome set". *Trials*. 2014;15:247. <https://doi.org/10.1186/1745-6215-15-247>.
9. Williamson P, Altman DG, Bagley H, Barnes KL, Blazeby JM, Brookes ST, Clarke M, Gargon E, Gorst S, Harman N, Kirkham JJ, McNair A, Prinsen CAC, Schmitt J, Terwee CB, Young B. The COMET handbook: version 1.0. *Trials*. 2017;18:280. <https://doi.org/10.1186/s13063-017-1978-4>.
10. Waltho D, Kaur MN, Haynes RB, Farrokhyar F, Thoma A. Users' guide to the surgical literature: how to perform a high-quality literature search. *Can J Surg*. 2015;58:348–58. <https://doi.org/10.1503/cjs.017314>.
11. Benstoem C, Moza A, Meybohm P, Stoppe C. A core outcome set for adult cardiac surgery trials: a consensus study. *PLoS One*. 2017;12:e0186772. <https://doi.org/10.1371/journal.pone.0186772>.
12. Benstoem C, Moza A, Autschbach R, Stoppe C, Goetzenich A. Evaluating outcomes used in cardiothoracic surgery interventional research: a systematic review of reviews to develop a core outcome set. *PLoS One*. 2015;10:e0122204. <https://doi.org/10.1371/journal.pone.0122204>.
13. Garg A, Rao S, Visveswaran G, Agrawal S, Sharma A, Garg L, Mahata I, Garg J, Singal D, Cohen M, Kostis JB. Transcatheter aortic valve replacement versus surgical valve replacement in low-intermediate surgical risk patients: a systematic review and meta-analysis. *J Invasive Cardiol*. 2017;29:209–16.
14. Schmidt T, Abbott J. Coronary stents: history, design, and construction. *J Clin Med*. 2018;7:126. <https://doi.org/10.3390/jcm7060126>.
15. Hall DA, Smith H, Heffernan E, Fackrell K. Core outcome measures in tinnitus international Delphi (COMiT-ID) research steering group. Recruiting and retaining participants in e-Delphi surveys for core outcome set development: evaluating the COMiT-ID study. *PLoS One*. 2018;13(7):e0201378. <https://doi.org/10.1371/journal.pone.0201378>.
16. Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino MA, Conaghan PG, Bingham CO 3rd, Brooks P, Landewé R, March L, Simon LS, Singh JA, Strand V, Tugwell P. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. *J Clin Epidemiol*. 2014;67:745–53. <https://doi.org/10.1016/j.jclinepi.2013.11.013>.
17. Myles P. Meaningful outcome measures in cardiac surgery. *J Extra Corpo Technol*. 2014;46:23–7. <https://doi.org/10.1016/j.jclinepi.2013.11.013>.
18. Swift SL, Puehler T, Misso K, Lang SH, Forbes C, Kleijnen J, Danner M, Kuhn C, Haneya A, Seoudy H, Cremer J, Frey N, Lutter G, Wolff R, Scheibler F, Wehkamp K, Frank D. Transcatheter aortic valve implantation versus surgical aortic valve replacement in patients with severe aortic stenosis: a systematic review and meta-analysis. *BMJ Open*. 2021;11(12):e054222. <https://doi.org/10.1136/bmjopen-2021-054222>.

Index

A

Absolute risk reduction, 120
ACR-TIRADS, 243
Action Research Arm Test, 71
Actual post-hoc effect size, 349
AGREE-S, 372–375, 378
AIDS, 196
Akaike Information Criterion (AIC), 85
American Joint Committee on Cancer (AJCC), 295
American Society of Anesthesiologists (ASA), 295, 313
AMSTAR 2, 70
Analysis of variance (ANOVA), 62, 222
Aortic insufficiency, 255
Aortic stenosis, 257
Aortic Valve Insufficiency and ascending aorta
 Aneurysm InternATIOnal Registry
 (AVIATOR), 258
Appraisal, 36, 37, 40, 47, 57, 58, 104, 152
Appraisal of Guidelines, Research and Evaluation
 (AGREE)-II, 231, 371
 checklist, 402
 tool, 231, 371
Appraise, 52, 152, 244, 294
Arthroscopic partial meniscectomy (APM), 139, 140,
 146, 147
Artificial intelligence (AI), 72
Atherosclerotic disease, 359
Attrition, 59

B

Background questions, 22
Bayesian Information Criterion (BIC), 85
Benefits, 50, 118, 152, 159–160, 271, 292, 294, 297, 298
Berlin Questionnaire, 5
Bethesda system, 243
Bias, 37, 44, 46, 107, 114, 116–120, 155, 156, 160, 191,
 192, 196, 197, 199, 201, 202, 269, 274, 293,
 295, 298, 306, 309, 327
Binary vs. continuous outcomes, 61
Biochemical markers, 65

Bisphosphonate, 99

Blinding, 44, 47, 49, 114, 115, 117, 156, 197

Body mass index (BMI), 65, 263, 295, 347

BOLERO-2 study, 98

Bone mineral density (BMD) and fracture outcomes,
 99, 100

Boolean operators, 28, 30, 31, 40

Brachial plexus injuries (BPI), 301

Breast implant-associated anaplastic large cell
 lymphomas (BI-ALCL), 222

Breast reconstruction, 178

British NHS, 284

Bubble-plots, 390

C

Cancer risk classification, 236

Capsular contracture, 178

Capture-mark-recapture (CMR) techniques, 180

Case-control studies, 46, 49, 50, 191, 206, 209, 257, 292,
 293, 295

 appraisal of selected article, 207

 circumstances and methods, 208

 clinical practice

 appraised study, 212

 benefits, 212

 follow-up, 211

 magnitude of risk, 211, 212

 occurrence, 211

 dose-response relationship, 208

 exposure and outcome, 209

 hazardous effect, 210

 indication/circumstances, 208

 odds ratio and confidence interval, 212

 PICOT format, 207

 reader, information for, 213, 214

 resolution of scenario, 212

 risk estimation, 210, 211

 temporal relationship, 208

 unmatched, 210

Case report forms (CRFs), 167

- Case series, 46, 49, 50, 217
 - appraisal of, 219
 - breast surgery, 217
 - change practice, 223
 - characteristics, 218–219
 - evidence-based medicine, 223
 - intervention, 221
 - literature search, 218
 - objective, 219
 - outcome assessment, 220–222
 - prospective study, 219, 220
 - recruitment consecutive, 220
- Causal indicators, 78
- Ceiling effects, 85
- CHEERS statement, 289
- Chi-square test, 153, 185, 316, 317
- Chlorhexidine, 385, 390, 391, 394
- Cholecystitis, 192
- Chylothorax, 206, 208, 209, 211, 212
- Circumferential resection margin (CRM), 337–338
- Classical test theory (CTT), 82–84
- Clinical decision-making, 55
- Clinical experimentation, 21
- Clinical practice guidelines (CPGs), 32, 35–37, 231, 232, 371
- Clinical question, 27–29, 35, 43, 52, 153, 158, 193, 201, 202, 273
- Clinical relevance, 122
- Clinical research question, 22, 104
 - background questions, 22
 - FINER criteria, 22, 23
 - foreground questions, 22
 - PICOT format, 24–26
- Clinical scenario, 27, 151
- Clinical sensibility testing, 306, 307
- Clinical trials, 114, 153, 205
- Clinician-based outcome measures (CBOMs), 70
- Cluster-randomized trials
 - allocation performance, 171, 172
 - challenges and risks of bias, 168, 169
 - cluster identification and selection, 166, 167
 - cluster-randomized crossover trials, 166
 - contamination, 165
 - crossover, 166
 - ethical considerations in, 170
 - evaluation, 171
 - intervention of interest, 164
 - parallel design, 165
 - pilot, 167, 168
 - pragmatic/explanatory, 164
 - randomization, 168
 - reporting, 168
 - sample size calculation, 172
 - statistical considerations in, 169
 - analysis, 170
 - ICC, 169
 - sample sizes, 169, 170
 - unit of inference, 169
 - stepped wedge design, 165, 166
- COCHRANE, 282
- Cochrane collaboration, 5
- Cohort studies, 47, 49, 52, 191–193, 195, 197, 198, 202, 203, 206, 293
- Colorectal cancer, 342
- COMET handbook, 409, 410
- Comparative intervention, 24
- Complex search strategy, 32
- Computer-generated randomization module, 59
- Conceptual framework, 78, 79, 126
- Confidence intervals (CIs), 121, 186, 199, 211, 266, 269, 315, 321, 329, 330
 - article validity and applicability, 337
 - calculation, 332
 - Cochrane systematic review and meta-analysis, 330
 - factors, 332
 - hypothesis testing, 332, 333
 - laparoscopic surgery, 337
 - minimally important difference and GRADE, 334–336
 - non-inferiority trials, 337, 338
 - PICOT, 330
 - power and sample size, 350
 - p-values, 336
 - SSI and mortality, 331
 - standard deviation and standard error, 331
 - systematic errors and bias, 331
 - systematic review and meta-analysis, 330, 336
 - univariate and multivariable model, 212
 - use of, 336
- Confirmatory factor analysis (CFA), 84
- Conflicts of interest (COI), 402
- Confounding, 59, 164, 196–199, 202, 293
- Consensus-based Standards for the selection of health Measurement Instruments (COSMIN), 6, 70, 87, 88
- Consistency/coherence checks, 391
 - evaluation for
 - global incoherence, 392
 - local incoherence, 391
 - qualitative check for, 391
 - quantitative check for, 391
- Consolidated Standards of Reporting Trials (CONSORT), 60, 127, 130–134, 166, 168
- Content analysis, 80
- Content validity, 81
- Continuous process improvement (CPI), 229
- Controlled vocabulary, 29, 30, 32, 34
- Core Outcome Measures in Effectiveness Trials (COMET), 7, 408
- Core outcome sets (COS), 187
 - COMET, 408, 411
 - guides
 - authors reported and adhered to protocol, 410
 - cerebrovascular complications, 411
 - multiple sources, 410
 - need for, 409, 410
 - outcomes, 410, 411
 - potential conflicts of, 411
 - scope of, 409
 - stakeholders, 410

- RCT, 408
 - SAVR-specific outcomes, 411
 - surgical specialties, 408
 - systematic reviews and meta-analyses, 407
 - users' guide, 408
 - Core Outcome Set-STAndards for Development (COS-STAD), 409
 - Cost analysis (CA), 159, 278, 280
 - Cost-benefit analysis (CBA), 278, 281
 - Cost-effective, 159
 - Cost effectiveness, 288
 - Cost-effectiveness analysis (CEA), 109, 278, 281
 - Cost-utility analysis (CUA), 278, 281
 - COVID-19 pandemic, 69
 - COX-1 selective NSAIDs, 297
 - Creating a search strategy, 31
- D**
- Data quality, 220
 - Decision analysis, 263–266, 270, 272–274
 - baseline analysis, 271, 272
 - caring
 - clinical features, probability estimates, 273, 274
 - generalization, 273
 - components of
 - clinical application, 266
 - decision tree, 264–266
 - outcomes, 266
 - users' guide, 266, 267
 - evidence change, uncertainty in, 272, 273
 - explicit and sensible process, 269, 270
 - literature search, 264
 - strategies and outcomes, 268, 269
 - in surgery, 264
 - uncertainty, potential impact of, 270, 271
 - utilities, 270
 - Decision analytic model, 268
 - Decision-making, 35, 36, 221, 232, 270
 - Decision pathway, 266
 - Decision tree, 50, 264–266, 269, 271, 272
 - Deep inferior epigastric perforator (DIEP) flap, 22, 279
 - Deep vein thrombosis (DVT), 292
 - Deliberate nonprobability-based sampling, 304
 - Delphi process, 401
 - Delphi technique, 305
 - Denosumab, 99
 - Descriptive cross-sectional survey, 309
 - Design effect (DE), 169
 - Deterministic analysis, 278
 - Developing A Surgical/Clinical question, 27
 - Diagnostic research, 47
 - Diagnostic studies, 235
 - ACR-TIRADS, 243
 - aspects of, 238
 - Bethesda system, 243
 - clinical scenario, 236
 - detection bias, 236
 - diagnostic tests, terms for, 239
 - likelihood ratio, 238, 240, 241
 - literature search, 236, 237
 - malignancy, risk of, 243
 - negative history and non-palpable nodule, 236
 - PICOT format, 236
 - positive predictive value and negative predictive value, 242
 - reference standard, 244, 245
 - likelihood ratios, 245, 246
 - management, change, 247
 - pre-test probability, 246
 - reproducibility of test, 246
 - rule in, rule out line, 247
 - sequential tests, 248
 - TIRADS, 245
 - methods, 245
 - US-guided FNAB, 246
 - selection bias, 237
 - sensitivity and specificity, 241, 242
 - surgery averse, 236
 - thyroid US, 237
 - TIRADS, 236, 237, 241
 - US ACR- and K-TIRADS, 237
 - US and FNAB for thyroid nodules, 237
 - US-guided FNAB, 243
 - Diagnostic tests, 44, 46, 47, 49, 236, 241–244, 247
 - Disutilities, 266
 - Dose-response gradient, 296
 - Dose-response relationship, 208
 - Duke Clinical Research Institute (DCRI), 4
- E**
- Economic analysis, 108
 - Economic evaluation, 50, 109, 266, 277, 278, 283
 - among subgroups, 287
 - caring
 - costs, 289
 - health outcomes, clinician's patient, 289
 - treatment benefits, 288, 289
 - CHEERS statement, 289
 - clinical strategies, 283, 284
 - cost-effectiveness, 283
 - costs and outcomes
 - clinical effectiveness, 284
 - costs measurement, 284, 285
 - integration, 285
 - treatment population, baseline risk in, 286
 - uncertainties, appropriate allowance, 285, 286
 - formulas
 - cost-analysis, 280
 - cost-benefit analysis, 281
 - cost-effectiveness analysis, 281
 - cost-utility analysis, 281
 - lumbar spinal stenosis, 282, 283
 - sensitivity analysis, 282
 - health care strategies, 283
 - incremental costs and outcomes, 286
 - intention, 286
 - National Health Service, 283
 - types, 278–280
 - uncertainty change, allowance for, 287, 288

- Economic research, 50
 - Effect indicators, 78
 - Effective sample size (ESS), 169
 - Enhanced recovery after surgery (ERAS), 397
 - appraisal
 - objective and health question of, 399
 - population, to guideline, 399
 - scope and purpose of, 399
 - for bariatric surgery, 398
 - development of, 398
 - guideline development, standardization of, 403
 - guidelines, appraisal, 398
 - literature search, 398
 - methodology and standardization,
 - recommendations for
 - elicited components of, 401
 - guideline development group with appropriate and stakeholders, 400
 - literature search, 400
 - multimodal and interprofessional elements, 401
 - quality of evidence, 400, 401
 - organization, 402, 403
 - program, 291
 - protocol, 165
 - reporting elements
 - identification, recommendations, 401, 402
 - peer-reviewed journal and ERAS® Society website, 402
 - potential conflicts of interest, 402
 - search strategy, 403
 - Enhancing the QUALity and Transparency Of health Research (EQUATOR) Network,
 - 6–8, 289, 398
 - Ergonomic/human engineering, 231
 - ESCAPE trial, 141–143, 145, 146, 148
 - Estimation, 196
 - Estrogen replacement therapy (ERT), 97
 - Ethical issues, 205
 - EuroQOL 5-Dimension (EQ-5D), 280
 - Evaluation, 35, 36, 43, 44, 104, 109, 156, 158, 201, 230, 231, 246, 307
 - Everolimus, 98
 - Evidence, 27, 28, 30, 31, 35–38, 40, 43, 44, 46, 49, 52, 110, 113, 114, 122, 153, 154, 159, 202, 203, 231, 264, 268, 270, 272, 292, 293, 298
 - Evidence-based medicine, 2, 11, 28, 31, 32, 35–37, 43, 113, 213, 228, 231
 - Evidence-based resources, 35
 - Evidence-based surgery (EBS), 2, 11, 35, 43, 52, 104, 160, 191, 292
 - appraise, 13–15, 18, 19
 - clinical practice, 13, 19
 - definition of, 4, 11
 - effectiveness and efficiency, 13, 19
 - history, 2, 5
 - COSMIN initiative, 6
 - CTFPHE's Examination's levels of evidence, 2, 3
 - EQUATOR Network, 6, 7
 - OMERACT, 5
 - PROSPERO, 7
 - relevant, answerable question construction, 12, 14
 - search of literature, 12–16, 18
 - steps, 12
 - Evidence synthesis, 70
 - Exclusion criteria, 64
 - Expertise, 115, 122, 152–157, 160, 243
 - Expert opinion, 46, 231
 - Explanatory trials, 164
 - Exploratory factor analysis (EFA), 84
 - External validity, 64, 294
- F**
- Feasibility investigation, 126
 - Feasibility studies, 167
 - Feasibility trial, 126
 - Fine-needle aspiration biopsy (FNAB), 235
 - FINER criteria, 22, 23
 - Fisher Exact test, 320
 - 5-point Likert scale, 75
 - Fixed-margin approach, 140
 - Floor effects, 85
 - Flosequinan, 97
 - Focus groups, 78
 - Foreground questions, 22
 - Forest plots, 390
 - FRESNO, 5
 - Fugl-Meyer assessment, 71
- G**
- Gastroesophageal reflux disease (GERD), 372
 - Gemini, 358
 - Generalizability, 121, 310
 - Generalization, 344
 - Generalized linear mixed models (GLMM), 170
 - Generalizing estimating equations (GEE), 170
 - Global incoherence, 392
 - Google and Google Scholar, 37
 - GRADEPro, 377
 - Grading of Recommendations Assessment, Development and Evaluation (GRADE), 5, 334–336, 377, 392, 400
 - Graphical Appraisal Tool for Epidemiological studies (GATE) framework, 213
 - Guide, 35, 104, 153
 - Guideline development group (GDG), 400
- H**
- Harm, studies reporting, 292, 293
 - patients/clinical practice, 296, 297
 - benefits, 297, 298
 - number needed to harm, 297
 - occurrence, 297
 - PICOT, 293
 - study appraisal, 294

- circumstances and methods, 295
 - dose-response gradient, 296
 - exposure and outcome, 296
 - follow-up, 295
 - patient characteristics, 295
 - risk estimation, 296
 - study valid, 294, 295
 - temporal relationship, 295
- study designs, 293
- Hazardous effect, 210
- Healthcare, 230
- Healthcare Resource Group codes, 285
- Health care strategies, 283
- Health-related outcome measures, 76
- Health-related quality of life (HRQOL), 104, 105, 109, 111, 343–345
- Health Utilities Index, 280
- Hernia repair, 56
- Herniorrhaphy, 56
- Heterogeneity, 184
- Hierarchical environment, 2
- Hierarchy of evidence, 35, 43, 44, 46, 47, 50, 52, 231
- Highly reliable organizations (HROs), 230
- Hotelling's T Squared test, 323
- Hypergeometric distribution, 320
- Hypertension, 313
- Hypothesis testing, 332, 333

I

- Idea, Development, Exploration, Assessment, and Long-term Follow-up (IDEAL) framework, 128, 131
- Identifying concepts, 28
- Identifying search terms, 28
- Importance of the question, 28
- Imprecision, 393
- Inception cohort, 220
- Inconsistency, 392
- Incremental cost effectiveness ratio (ICER), 281, 285
- Indeterminate nodules, 245
- Indirectness, 392
- Information bias, 197
- Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT), 147
- Institute of Medicine (IOM), 229
- Institutional review boards (IRBs), 170
- Instrumental variables, 200
- Intention-to-treat (ITT)
 - analysis, 118, 144, 317, 324
 - modification, 322
 - principle, 60
- Internal consistency, 85
- Internal validity, 58, 66, 294
- International Knee Documentation Committee (IKDC) questionnaire, 145

- International Prospective Register of Systematic Reviews (PROSPERO) database, 182

- Interpret scores, 85
- Interspinous distractor device (IDD), 282
- Intervention, 28, 30, 34, 44, 46, 47, 50, 104, 105, 108–110, 114, 115, 117, 119, 121, 122, 151–154, 156, 159, 160, 193, 195, 236, 266, 268, 273, 274, 292, 293
- Intra-cluster correlation (ICC), 168, 169
- Iodine povacrylex, 170, 172
- Item Response Theory (IRT), 82–84

K

- Keeping current, 38
- Keywords, 28–30, 34, 105, 153

L

- Laminectomy, 282, 286
- Laparoscopic surgery, 56
- Lasofloxifene, 99
- League tables, 388–390
- Length of hospital stay (LOS), 397
- Levels of evidence (LOE), 2, 3, 43, 44, 46, 47, 49, 50, 52–54, 56, 153, 160
- Libra, 358
- Likelihood ratio (LR), 47, 238–248
- Limits, 30, 32, 34, 40, 114, 116, 157, 230
- Lind, James, 2
- Literature appraisal, 104
- Literature search, 104, 114, 153, 293
- Local incoherence, 391–392
- Loss to follow-up (LTFU), 222
- Lumbar spinal stenosis (LSS), 282, 283

M

- Malignant polyp, 313
- Marfan syndrome, 256
- Mastectomy, 178
- Mechanical bowel preparation (MBP), 330
- Mechanical valve thrombosis, 259
- Medical Subject Headings (MeSH), 56
- MEDLINE database, 179
- Meta-analysis, 177, 178, 183
- Minimally important difference (MID), 89, 109, 145, 334–336
- Minimally invasive surgery, 56, 65, 66
- Minimally invasive techniques, 284
- Minimum important difference (MID), 318
- Minitab, 320, 322
- Modern psychometric approach, 85
- Motor vehicle collision (MVC), 358
- Multidomain conceptual frameworks, 78
- Multiplicity, 323
- Multivariable analyses, 62
- Multivariable model, 210
- Multivariable regression, 62, 63

N

Narrative reviews, 178
 National Health Services (NHS), 230
 National Surgical Quality Improvement Program (NSQIP), 228
 Needed to treat (NNT), 63
 Negative pressure wound therapy (NPWT), 128
 Netranking P-scores, 389
 Network graphs, 387, 388
 Network meta-analyses (NMA), 383
 clinical question, 387
 confidence, evaluation, 392
 imprecision, 393
 inconsistency, 392
 indirectness, 392
 publication bias, 393
 study limitations, 392
 confidence interval, 390
 consistency/coherence checks, 391
 global incoherence, 392
 local incoherence, 391
 qualitative check for, 391
 quantitative check for, 391
 elements of, 384
 evaluation, 385
 forest plots, 390
 fundamentals
 comparisons in, 386
 inclusion criteria for, 385, 386
 high methodologic quality, primary studies of, 387
 league tables and ranking metrics, 388–390
 network graphs, 387, 388
 patient care, 393, 394
 costs and potential risks, 394
 outcomes, 394
 return to case scenario, 394
 PICO, 384
 SSI, 384
 statistical analyses, 388
 studies detailed and exhaustive search, 387
 studies reproducible, assessments of, 387
 study to study, 390
 systematic review, 385
 vs. traditional pairwise meta-analyses, 384
 Nissen, 378
 Non-inferiority hypothesis, 359
 Non-inferiority margin, 140, 145
 Non-inferiority randomized controlled trials,
 140, 141, 147
 confidence intervals, 337, 338
 critical appraisal framework, 141
 ESCAPE trial, 142, 143, 146
 margin, 145
 novel and standard surgical intervention groups, 142
 participants, 142, 144, 145
 patient-important outcomes, 147
 PT, 148
 test, 145, 146
 Non pre-appraised research, 36
 Non-steroidal anti-inflammatory drugs (NSAIDs),
 292, 293, 296–298

Null hypothesis, 317, 345

Number needed to harm, 297

Number needed to treat, 120, 121, 297

Numeric rating scale (NRS), 343, 348

O

Observational studies, 44, 46, 49, 191, 199, 292
 Odds ratio (OR), 62, 63, 199, 211, 212, 293, 296
 Olanexidine, 385, 390, 394
 OMERACT, 5
 One-on-one interviews, 78
 Open reduction internal fixation (ORIF), 358
 Oral antibiotics (oAB), 330
 Organization for Economic Cooperation
 and Development (OECD), 278
 Orthopaedic surgery, 263
 Outcome detection, 206
 Outcome measures (OMs), 72, 109, 122, 293
 critical appraisal of evidence, 69
 impacts, 70
 search and critical appraisal, 70–72
 Outcomes, 28, 44, 46, 47, 49, 50, 103–105, 107, 108,
 114–117, 119, 122, 154–156, 158, 159,
 191–193, 195–197, 201–203, 228–231, 236,
 271, 292–294, 296

P

Parallel cluster-randomized trial, 165
 Partial economic evaluation, 278
 Partial fundoplication, 378
 PASS software, 346
 Patient and Observer Scar Assessment Scale
 (POSAS), 89
 Patient-reported outcome measures (PROMs), 69
 analyze field test data, 84, 85
 application, 76
 components, 77
 conceptual framework, 78, 79
 construct, 77
 content validity of, 81
 CTT, 82, 83
 development, 77
 field test of, 81, 82
 first draft, develop and select items for, 79–81
 IRT, 83, 84
 RMT, 83, 84
 SCAR-Q, 89, 90
 selection, 88, 89
 Patient-reported outcomes (PRO), 76, 105, 106, 109, 158
 Patient safety, 228
 Patient Scar Assessment Questionnaire (PSAQ), 90
 Pearson method, 318
 Peer-reviewed journal, 402
 Peg Insertion Test, 71
 Performance-based function tests, 72
 Performance-based outcome measures (PBOMs), 69–70
 Peripheral nerve injuries, 301
 Physical therapy (PT), 139, 148
 Pilot cluster-randomized trials, 167–168

- Pilot trials, 133–135
 - criteria of success, 130
 - current practice in, 127, 128
 - definition of, 126
 - methods of analysis, 129, 130
 - necessity of, 126, 127
 - objectives and outcomes, 128, 129
 - reporting, 130, 131
 - resources, 131, 132
 - sample size, 129
 - study rationale, 128
- Pilot work, 126
- POISE trial, 58
- Polish National Health Services, 258
- Polish National Registry of Cardiac Surgery Procedures, 258
- Population, intervention, comparison intervention, time horizon (PICOT), 12, 104, 105, 153, 236, 293
 - format, 24–26, 57, 171, 282, 358
 - question, 64
- Population, intervention, control, and outcome (PICO), 140
- Post anaesthetic recovery (PAR), 347
- Postoperative days (PODs), 348
- Post-operative nausea and vomiting (PONV), 401
- Postprandial abdominal pain, 192
- Post-test probability, 240, 241, 246, 248
- Power and sample size, 119, 120
 - appraisal, questions used to, 348
 - clinical experience and published trials, precise and consistent with, 350
 - confidence intervals, 350
 - effect size, 349, 350
 - estimation, 350
 - mean pain scores, 349
 - pain scores, 349
 - performance, 348
 - primary outcome, calculation for, 348, 349
 - treatment effect, 350
 - appraised article, summary of, 347, 348
 - article selection, 346
 - importance of, 342–346
 - pain outcomes, 347
 - PICOT format, 346, 347
- Pragmatic trials, 164
- Preappraised research, 35
- Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA), 182, 183
- Prehabilitation, 223
- Preoperative Aqueous Antiseptic Skin Solution in Open Fractures (Aqueous-PREP), 127
- PREPARE trial, 171, 172
- Pre-test probability, 240, 241, 246, 248
- Probabilistic analysis, 278, 282
- Prognosis, 28, 34, 35, 46, 117, 203
- Prognostic research, 49
- Prognostic study
 - aortic stenosis, 257
 - appraising surgical literature on, 256, 258
 - follow-up, 258, 259
 - likelihood, estimates of, 259, 260
 - objective and unbiased outcome criteria, 259
 - outcomes, 259
 - patient characteristics and demographics, 260
 - sample representative, 258
 - clinical scenario, resolution of, 260
 - physical examination, 255
 - study designs for, 256, 257
 - surgical literature, user's guide to, 257
- Prospective cohort studies, 40, 46, 49–51, 191, 192, 293, 295
 - analytical approach, 199
 - appraisal, practical guide for, 192
 - bias, potential sources of, 196
 - cholecystitis, 192
 - clinical epidemiology, 193, 194
 - clinical practice
 - broader context of clinical question, study placed in, 201
 - existing literature, description of, 201, 202
 - outcomes, 202
 - clinical scenario, resolution of, 202
 - confounding, 197, 198
 - data collection methods, 195
 - data sources, 194, 195
 - exposure groups, characteristics of, 199
 - information bias, 197
 - limitations, 202
 - literature search, 192
 - precision of effect estimates, 199, 200
 - primary and secondary outcomes, 195, 196
 - research question and clinical relevance, 193
 - robot-assisted and laparoscopic cholecystectomy, 192
 - robust to planned sensitivity analyses, 200, 201
 - selection bias, 196, 197
 - statistical analysis, 198, 199
 - study element, 194
 - study population, 193, 194
 - subgroup analyses, 201
 - univariable analyses, 199
- Prospective Register of Systematic Reviews (PROSPERO), 7, 71
- Prostatectomy, 25
- Protective effect, 209
- Publication bias, 181, 393
- P values, 121, 199, 200, 314, 316–318
 - CIs from, 336
 - clinical scenario, resolution of, 323, 324
 - confidence intervals, 321
 - multiplicity, 323
 - outcome, 315
 - PICOT format, 314
 - randomized controlled trial, 314, 315
 - secondary outcome, 320
 - sensitivity analysis, 322, 323
 - software, 321, 322
 - statistical principles, 314
 - statistical tests, 319–321
 - surgical literature, statistical findings in, 324
- Pyramid of EBM resources, 35

Q

Qualitative interviews, 78
 Quality Adjusted Life Years (QALYs), 266, 280, 283, 284
 Quality Appraisal for Clinical Measurement Research Reports Evaluation, 70
 Quality improvement (QI), 229, 230
 Quality improvement and patient safety (QIPS), 227, 228
 Canadian provincial database, 231
 clinical practice guidelines, 231, 232
 clinical scenario, resolution of, 232
 ergonomic/human engineering, 231
 healthcare, 230
 IOM, 229
 NSQIP, 229
 patient demographics and outcomes, 228
 progressive change, 228
 research, 229, 230
 shared decision model, 228
 surgery, 228, 230, 231
 Swiss Cheese Model, 231
 Quality of life (QoL), 22, 104, 106–109, 151, 153, 156, 158–160, 266, 270, 271, 273, 292
 Questionnaire, 105–109, 305–307

R

RAND-36 Physical Component Score, 147
 Randomization, 44, 59, 60, 114–117, 152, 154–157, 168, 293
 Randomized controlled trials (RCTs), 34, 44–47, 49, 50, 104, 105, 111, 114, 115, 117–119, 125, 130, 140, 151–157, 159, 160, 191, 192, 197, 203, 205, 264, 270, 272, 274, 292, 293, 301, 341, 385, 408
 Range of motion (ROM), 263
 Ranking metrics, 388, 389
 Rankograms, 390
 Rasch measurement theory (RMT), 82–85, 89
 RedCAP database, 306
 Reference distributions, 317
 Reflective framework, 78
 Rehospitalization, 412
 Relative risk (RR), 199, 293, 320, 330
 Reliability, 107, 108, 195, 302, 307, 308
 Reporting on ERAS Compliance, Outcomes, and Elements Research (RECOVER) Checklist, 398, 402–403
 Research, 27, 28, 36–38, 40, 44, 46, 49, 50, 52, 104, 106, 113, 115, 117, 156, 158, 160, 193, 194, 229–231, 238, 294, 306
 question, 27, 36, 43, 44, 47, 49, 105, 114, 152, 153, 236, 293, 304
 Resolution of the clinical scenario, 40, 122, 160
 Response rate, 308
 Retrospective cohort studies, 46, 49, 50, 59, 191, 192, 194, 295
 analytical approach, 199
 appraisal, practical guide for, 192

 bias, potential sources of, 196
 broader context of clinical question, study placed in, 201
 cholecystitis, 192
 clinical epidemiology, 193, 194
 clinical practice
 existing literature, description of, 201, 202
 outcomes, 202
 clinical scenario, resolution of, 202
 confounding, 197, 198
 data collection methods, 195
 data sources, 194, 195
 exposure groups, characteristics of, 199
 information bias, 197
 limitations, 202
 literature search, 192
 precision of effect estimates, 199, 200
 primary and secondary outcomes,
 definitions of, 195, 196
 research question and clinical relevance, 193
 robot-assisted and laparoscopic cholecystectomy, 192
 robust to planned sensitivity analyses, 200, 201
 selection bias, 196, 197
 statistical analysis, 198, 199
 study element, 194
 study population, 193, 194
 subgroup analyses, 201
 univariable analyses, 199
 Risedronate, 100
 Risk, 28, 46, 49, 104, 114, 120–122, 152, 159–160, 195, 229, 236, 271, 273, 292–298
 Risk adjustment for congenital heart surgery (RACHS-1) score, 209
 Risk of bias assessment, 14
 Risk Ratios (RR), 62, 63
 Robot-assisted approach, 192
 Robotic surgery, 56, 61, 227

S

SAGES guideline, 376
 Sample size, 44, 60, 96, 119, 120, 129, 153, 169, 170, 192, 193, 200, 292, 296
 Sampling element, 303
 Sampling frame, 303
 SCAR-Q, 89, 90
 Searching, 28, 29, 31, 32, 37, 40
 Search strategy, 28, 56, 70
 Selection bias, prospective and retrospective cohort studies, 196, 197
 Sensitivity analysis, 153, 154, 185, 264, 266, 269–274, 282, 322–323, 326–327
 Seoul National University Hospital Patients Research Environment (SUPREME) system, 294
 Separating Indirect from Direct Evidence (SIDE) approach, 391
 Simple randomization, 163
 Simple search strategy, 32
 6S pyramid, 36

- Smallest detectable change (SDC), 85
- SQUIRE, 229, 230
- Standard deviation (SD)
 - confidence intervals, 331
 - power and sample size, 345, 346
- Standard error (SE), 331
- Standard error of measurement (SEM), 89
- Stanford University has a Center for Clinical Research (SCCR), 4–5
- Statistical analysis, 192
 - NMA, 388
 - prospective and retrospective cohort studies, 198, 199
- Statistical inefficiency, 169
- Statistical inference, 344
- Statistical methods, surgical research, 308, 309
- Statistical tests, 199, 316–318
 - clinical scenario, resolution of, 323, 324
 - computing, 318
 - confidence intervals, 321
 - multiplicity, 323
 - number mean, 318, 319
 - outcome, 315
 - PICOT format, 314
 - P values, 319–321
 - randomized controlled trial, 314, 315
 - secondary outcome, 320
 - sensitivity analysis, 322, 323
 - software, 321, 322
 - statistical principles, 314
 - surgical literature, statistical findings in, 324
- Statistics, 193, 199
- Stepped wedge design, 165, 166
- Student's t distribution, 321
- Student's t test, 321
- Study design for therapeutic studies, 45
- Subgroup, 357, 358
 - appraised article, 359
 - composite bleeding outcome, 361
 - credibility, criteria to, 361
 - MVC, 358
 - noninferiority hypothesis, 359
 - nonorthopedic surgery subgroup, 361
 - PICOT format, 358
 - rationale indication, 362
 - TXA, 359
- Subgroup analysis, 119, 201
- Summaries and guidelines, 35
- Superior treatment, 65
- Suprafascial perforator flaps, 23
- Suprafascial peri-umbilical perforator flap, 23, 24
- Surgery, 27–30, 32, 34, 36, 40, 43, 103, 104, 108, 117, 152, 156, 159, 197, 228, 237, 247, 248, 264, 266, 292, 293, 295, 297, 298
- Surgical aortic valve replacement (SAVR), 408, 409
- Surgical guideline appraisal, 371, 372
 - AGREE-S, 372–375
 - development of recommendations and editorial independence, 377
 - evidence synthesis, 376, 377
- GERD, 372
 - implementation and update, 378
- PubMed, 372
- SAGES guidelines, 372
- scope and purpose, 373
- stakeholders, 373, 376
- Surgical interventions
 - appraisal, 57, 58
 - attrition, 59
 - binary vs. continuous outcomes, 61
 - blinding, 59
 - follow-up, 60
 - intention-to-treat principle, 60
 - odds ratios (OR), 62, 63
 - outcomes, 61, 65
 - patient history, 56
 - PICOT format, 57
 - search strategy, 56
 - surgical skills, 65
 - treatment effect, 62–64
 - univariable and multivariable analyses, 62
- Surgical proficiency, 65
- Surgical research, 44
- Surgical research methodology (SRM) course, 5
- Surgical research, surveys and questionnaires, 301
 - appraisal, 303
 - change practice, 310
 - characteristics, 302
 - clinical scenario, resolution of, 310
 - framework, 303
 - literature search, 302
 - peripheral nerve injuries, 301
 - primary guides
 - administration of, 306
 - appropriate development, 304–306
 - research question and objective, 303
 - sampling frame, appropriate selection, 303, 304
 - secondary guides
 - clinical sensibility testing, 306, 307
 - conclusions appropriate, 309, 310
 - pilot testing, 306
 - reliability and validity testing, 307, 308
 - reporting of, 309
 - response rate, 308
 - statistical methods, 308, 309
- Surgical seclusions, 4
- Surgical site infection (SSI), 384
- Surrogate endpoints, 95
 - advantages, 96
 - bone mineral density and fracture outcomes, 99, 100
 - clinical benefit, 96
 - limitations, 97, 98
 - validation, 98
- Survey, 109, 304, 306, 307, 310
- Synopses of single studies, 37
- Synopses of syntheses, 36
- Systematic deviation, 55

Systematic review (SR), 70, 71, 75, 177, 178
 and application, 187
 CI, 186
 clinical question, 179, 180
 clinical scenario, resolution of, 188
 of high methodological quality, 181–183
 limitations, 71
 nonrandomized data, 185
 patient-important outcomes, 187
 potential costs and risks, 187, 188
 precision, 186
 publication bias, 181
 qualitative analysis, 186
 strengths, 71
 studies and exhaustive, search for, 180
 studies reproducible, assessments of, 183
 study to study, 183–185
 users' guide for, 179

T

Target population, 303
 Tegner Activity Scale, 147
 Telerehabilitation, 71
 Temporal relationship, 208
 Test statistics, 317
 Thematic analysis, 80
 Therapeutic research, 44
 Threshold levels, 271
 Thyroid Imaging Reporting and Data System (TIRADS),
 236, 237, 240–243, 245, 246, 248
 Thyroid nodules, 49, 236, 237, 242, 247
 Time horizon, 25
 Toupet fundoplication, 378
 Toupet repair, 378
 Traditional Chinese medicine (TCM), 127
 Tranexamic acid (TXA), 358, 359
 Transcatheter aortic valve replacement (TAVR),
 257, 408, 409
 Transitivity, 386
 Transverse rectus abdominis myocutaneous (TRAM)
 flap, 279
 Transversus abdominis plane (TAP), 25, 342
 Treatment, 28, 44, 46, 49, 50, 108, 109, 115–118,
 120–122, 152–160, 197, 199, 264,
 266, 292, 297

effect, 114, 117, 119, 152, 157, 158, 297
 standardization, 119
 TreeAge Pro, 269
 Trial, 34, 36, 37, 52, 107, 115, 117–122,
 152–160, 203
 Trial-based economic evaluation, 279
 Type I error, 58, 200, 344
 Type II error, 58, 60, 344

U

Ultrasound (US), 235
 Unit of inference, 169
 Unit of randomization, 163
 Univariable analyses, 62
 Univariable model, 210
 Upper extremity lymphedema, 217
 US Panel on Cost-effectiveness, 283
 Utility, 106, 109, 160, 231, 264, 266, 270–273

V

Validity, 107, 108, 114, 119, 154, 156–158,
 196, 198, 200, 201, 274, 293,
 302, 306–308, 310
 Varimax, 84
 Ventral hernia, 56
 Virgin surgery, 214
 Virtual reality, 71
 Visual Analog Scale (VAS), 344
 Vitamin D, 100

W

Welsh t test, 321
 Willingness-to-pay (WTP) approach, 281
 Workers Compensation Benefit (WCB), 284
 World Health Organization (WHO), 5

X

X-Stop, 286

Z

Zygomaticomaxillary complex (ZMC) fracture, 358