Professional Practice in Governance and Public Organizations

Fabio Berton
Paolo Paruolo *Editors*

Data-Driven Learning in the EU

From Administrative Sources to Public Value



OPEN ACCESS

Springer

Professional Practice in Governance and Public Organizations

The *Professional Practice in Governance and Public Organizations* series features cutting-edge insights and practical guidance for professionals in the areas of economics, politics, public policy, public administration, and international organizations. With concise, accessible volumes, the series explores the latest developments in governance, public law-making, organizational and political strategies, institutional policies, policy instruments, public management, and finance. Core topics include decision-making, leadership, and the transformative impact of digitalization.

Each book is authored by practitioners, experts, and leading authorities from public and international organizations, think tanks, and NGOs, ensuring a blend of theoretical depth and real-world applicability. While designed primarily for professionals in these fields, the series also serves as an invaluable resource for students of economics, political science, public policy, and public administration, equipping them with practical guides for their future careers.

Fabio Berton · Paolo Paruolo
Editors

Data-Driven Learning in the EU

From Administrative Sources to Public Value

Springer

Editors
Fabio Berton
European Commission Joint Research Centre
Ispra, Italy

Paolo Paruolo
European Commission Joint Research Centre
Ispra, Italy



ISSN 2731-9776 ISSN 2731-9784 (electronic)
Professional Practice in Governance and Public Organizations
ISBN 978-3-031-99031-1 ISBN 978-3-031-99032-8 (eBook)
https://doi.org/10.1007/978-3-031-99032-8

This work was supported by Joint Research Centre.

© European Commission Joint Research Centre 2026. This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.
The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
Unless otherwise expressly agreed by the Publisher in writing, the Publisher does not authorize use or reproduction of all or any part of this work in any manner for the purpose of training, fine-tuning, evaluating, or developing artificial intelligence models, technologies, or systems. In accordance with Article 4(3) of the EU Directive 2019/790 on copyright and related rights in the Digital Single Market, the Publisher expressly reserves all rights with respect to this work under the text and data mining exception of the directive.
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.

Fabio Berton dedicates this book to Jole and Michel
Paolo Paruolo dedicates this book to his family

Preface

This book is an ideal continuation of Crato and Paruolo's (eds., 2019) "Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design", Springer. It addresses two central questions: (1) What is the current capacity of the European Union to leverage administrative data for the purpose of learning which policies work, and for whom? and (2) What strategic enhancements could optimise the utilisation of such data to strengthen evidence-based policymaking?

Causal evidence of policy-effectiveness involves data-driven learning, defined as the process of learning (if and) how policies work and for whom, on the basis of observational data.[1] Administrative data refers to information collected for administrative purposes primarily, often by government agencies, organisations or institutions. This type of data is typically gathered during routine operations, such as managing programmes, services or resources, and it is generated as a byproduct of administrative processes.

This book's major messages are:

(i) Learning what works, for whom, and possibly why, has public value;
(ii) Data are a key ingredient in the production of scientific evidence about what works;
(iii) Knowledge arising from improved access to data is beneficial for all;
(iv) Many barriers, often arising from the asymmetric information about the benefits from improved data access, still exist;
(v) Administrative microdata occupy a privileged position to feed policy-relevant scientific knowledge about what works;
(vi) Monitoring using aggregate indicators provides different knowledge content than real-time evaluation;

[1] Understanding cause-and-effect relations is at the heart of science. It goes back to the classic Latin philosophical principle "Scientia est cognitio rerum per causas" (Science is the knowledge of things through their causes). The concept originates from Aristotle, and was later a cornerstone of Scholastic philosophy by St. Thomas Aquinas; it was then adopted in the modern Scientific Revolution in the transition to modern empirical science, and encapsulated in Francis Bacon's motto "Verum scire est per causas scire" (To know truly is to know through causes).

(vii) Artificial Intelligence and machine learning can help but require human supervision in data analysis and in supporting the creation of causal evidence;
(viii) The EU legal framework, aiming to protect data privacy, does not represent an unsourmountable hurdle to data access;
(ix) Administrative data can also be used proficiently to reduce the Member States' reporting burden to the European Commission;
(x) A persisting lack of support and coordination at European level may induce a decline of best practice in data access and this should be avoided.

This book has been a pilot experiment towards a fully-fledged handbook. It involved collecting evidence from 27 European countries that can guide (a) researchers through data-access channels, (b) data-holding institutions towards data-security best practices, and (c) policymakers on the evolving body of causal evidence about what works in several policy fields. This exercise could happen with periodical updates, possibly with the funding and supervision of the European Union. Moreover, this book's network of seventy authors did not exist previously, and it is a valuable community for future science-for-policy cooperation.

Finally, the book provides an account of how all EU Member States have something to learn from the experience of others. This can give rise to an EU community of practice for institutions regarding administrative data access and research, in a fashion that could be recognised, coordinated and supported by the European Union.

This book would not have been possible without the kind availability of the seventy experts who generously and enthusiastically accepted the invitation to contribute to this project with the sole purpose of creating public value. They are listed after the table of contents and their short bios and contacts appear at the end of each chapter. The editors sincerely thank them all.

The book has also benefited from the experience of colleagues at the Competence Centre on Microeconomic Evaluation (CC-ME), established at the European Commission Joint Research Centre more than 10 years ago on May 16, 2016. Some of them have authored chapters in this book, and all of them have contributed to it with continuous feedback and discussion during its two-year development. In a sense, the book is the result of CC-ME's collective experience in policy analysis and support.[2]

Similar discussions took place with other colleagues in national and international venues in which this book project was discussed and presented: the Third Workshop of the LABORatorio R. Revelli in Torino in 2023, when this book idea was first proposed following a roundtable on "Administrative data for Policy Evaluation"[3]; the 2024 COMPIE Conference in Amsterdam, in particular during the roundtable on cross-country research using administrative data[4]; the 2024 ECPR Conference in Dublin, with its session on the Better Regulation Agenda[5]; the NetCIEx workshop

[2] See: https://knowledge4policy.ec.europa.eu/microeconomic-evaluation_en.

[3] See: https://www.laboratoriorevelli.it/en.

[4] See https://knowledge4policy.ec.europa.eu/event/compie-2024-conference_en.

[5] See https://ecpr.eu/Events/Event/PanelDetails/14890.

held in Ispra in 2025.[6] The editors also wish to thank Sean O'Dubhghaill for his valuable proofreading, which allowed this book to benefit from a style homogeneity that is seldom reached in edited volumes.

Fabio Berton would also like to thank his mentor Bruno Contini, Professor Emeritus of Econometrics at the University of Torino, Italy, and a pioneer of the use of administrative data in Europe; he instilled the culture of data quality in him alongside colleagues from the LABORatorio Riccardo Revelli (who became close friends in the meantime). Paolo Paruolo would like to thank the JRC management and colleagues who supported the CC-ME's creation and development over the last 10 years. Finally, both editors want to thank the JRC for its support to this book project and their families for the countless hours that they sacrificed from their family lives for the editors' pursuit of the fleeting joy of knowledge.

Ispra, Italy
February 2026

Fabio Berton
Paolo Paruolo

[6] See https://knowledge4policy.ec.europa.eu/microeconomic-evaluation/netciex_en.

Contents

1 **Why This Book?** 1
Fabio Berton and Paolo Paruolo

Part I EU Experiences

2 **Working with Large Administrative Datasets from Austria** 21
Omar Bamieh and Lennart Ziegler

3 **The Use of Administrative Data for Evidence-Informed Policy Development in Belgium** 29
Koen Vleminckx

4 **On Some Aspects of Open Administrative Data in Bulgaria** 49
Nikolay Netov

5 **The Reality of Administrative Data in Croatia** 61
Nataša Nater Drvenkar, Alka Obadić, and Josip Tica

6 **Access and Use of Administrative Data: The Case of Cyprus** 77
Nikolaos Theodoropoulos and Andreas Tryphonides

7 **Learning from Administrative Data in Czechia: Hiccups at the Last Mile** 87
Petr Bouchal and Dominik Hajduk

8 **Data-Driven Research in Denmark** 107
Maria Schlier and Morten Bennedsen

9 **Estonian Administrative Data for Policy Evaluation** 117
Andres Võrk, Kadri Rootalu, and Jaan Masso

10 **Accessing Administrative Data for Research: Lessons from Finland** 137
Stefano Lombardi

11 Access to French Administrative Micro-Data 155
Brice Fabre and Kamel Gadouche

12 Quick and Easy, But Decentralised: Access to Administrative Data in Germany 165
Hanna Brenzel, Toni Schell, Tatjana Mika, and Stefanie Wolter

13 The Limited Use of Administrative Data in Greece 179
Daphne Nicolitsas

14 Rich Data, Strong Laws to Promote Access, Weaknesses and Uncertainties in Implementation: The Case of Hungary 189
Ágota Scharle

15 Building and Accessing a Work and Welfare Longitudinal Dataset from Irish Administrative Data 197
Hugh Cronin

16 The Long and Winding Road: The Access to Administrative Data for Research in Italy 213
Roberto Leombruni and Chiara Ardito

17 Administrative Data Access in Latvia 231
Ieva Bloma, Jānis Glazkovs, and Fabio Berton

18 Open Administrative Data in Lithuania 243
Egidijus Barcevičius

19 Administrative Data Access in Luxembourg 251
Michela Bia

20 Access to Administrative Data in Malta 267
Rose Marie Azzopardi

21 Statistics Netherlands Microdata: A User Perspective 281
Ivo Gorissen, Maarten Bloem, Vincent van Hees, and Ran van den Boom

22 Administrative Data for Research and Public Policy Evaluation in Poland: The Case of Science and Higher Education Sector 291
Kamila Lewandowska, Ewa Zegler-Poleska, and Emanuel Kulczycki

23 Administrative Data in Portugal 305
Francisco Lima, José Pinto Martins, Katja Neugebauer, Joana F. Pimentel, Sofia Rodrigues, Paulo Saraiva, Rita Sousa, and João Zilhão

24 Pathways to Accessing Administrative Microdata in Romania 327
Andreea Mitrut, Oana Borcan, and Alina Botezat

25 Slovak Administrative Data Increasing the Value for Money in Public Spending 333
Miroslav Štefánik

26 Unlocking Slovenia's Administrative Data: Abundant Sources, Complex Access 351
Matija Vodopivec

27 State of the Art in Access to Administrative Data in Spain 363
Josep Amer-Mestre and Manuel Serrano-Alarcón

28 Accessing Administrative Data for Research: Lessons from Sweden 377
Stefano Lombardi

Part II Horizontal Perspectives

29 Cross-Country Networks, Administrative Data and Policy Learning 395
Josep Amer-Mestre, Sara Flisi, and Paolo Paruolo

30 Harnessing Administrative Data for Policy Evaluation in Labour, Education and Health Economics 411
Nuno Crato, Cinzia Di Novi, and Elena Claudia Meroni

31 Learning About Environmental Outcomes in Europe 433
Zelda Brutti, Marzia Freo, and Laura Serlenga

32 Administrative Data Access on Firms and Markets 453
Stefano Verzillo, Tomaso Duso, Francesco Quatraro, Pietro Santoleri, and Alexander Schiersch

33 Current Challenges and Opportunities in Data-Driven Learning 483
Fabio Berton, Paolo Paruolo, and Ian Vollbracht

Editors and Contributors

About the Editors

Fabio Berton is Policy Analyst on impact evaluation at the Competence Centre for Microeconomic Evaluation at the Joint Research Centre of the European Commission. He is also Associate Professor of Economic Policy at the University of Torino, and Research Fellow at CIRET, IZA and LABORatorio R. Revelli. His research focuses on applied labour economics, and on the evaluation of labour policies in particular. Fabio has publications on *The Economic Journal, The British Journal of Industrial Relations, The Oxford Bulletin of Economics and Statistics, The Review of Financial Studies and The Journal of Economic Inequality* among the others.

Paolo Paruolo is an econometrician, currently Head of the European Commission Competence Centre on Microeconomic Evaluation since 2016. He holds a Ph.D. in econometrics from the University of Copenhagen, and he has been full professor of econometrics at the University of Insubria, Varese, Italy and previously at the University of Bologna. His research focuses on econometric theory, time series analysis and counterfactual methods. He has published in *Econometrica, Journal of Econometrics, Econometrics Journal, Econometric Theory, Journal of the Royal Statistical Society Series A, Oxford Bulletin of Economics and Statistics*, among others. He has written and edited several books, including *Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design*, Crato, Paruolo (eds., 2019).

Contributors

Josep Amer-Mestre European Commission Joint Research Centre, Ispra, Italy

Chiara Ardito European Commission Joint Research Centre, Ispra, Italy

Rose Marie Azzopardi University of Malta, Valletta, Malta

Omar Bamieh Andrés Bello University, Santiago, Chile

Egidijus Barcevičius PPMI Group, Vilnius, Lithuania

Morten Bennedsen University of Copenhagen, Copenhagen, Denmark

Fabio Berton European Commission Joint Research Centre, Ispra, Italy

Michela Bia LISER, Esch-Sur-Alzette, Luxembourg

Maarten Bloem CBS, The Hague, Netherlands

Ieva Bloma European Union Agency for Asylum, Valletta, Malta

Ran van den Boom CBS, The Hague, Netherlands

Oana Borcan University of East Anglia, Norwich, UK

Alina Botezat Gheorghe Zane Institute for Economic and Social Research, Romanian Academy, Iași, Romania

Petr Bouchal Government Analytical Unit, Government Office of the Czech Republic, Prague, Czech Republic

Hanna Brenzel Research Data Centre of the Federal Statistical Office, Wiesbaden, Germany

Zelda Brutti European Commission Joint Research Centre, Ispra, Italy

Nuno Crato University of Lisbon, Lisbon, Portugal

Hugh Cronin Department of Social Protection, Irish Government Economic and Evaluation Service (IGEES), Dublin, Ireland

Cinzia Di Novi University of Pavia, Pavia, Italy

Tomaso Duso German Institute for Economic Research (DIW), Berlin, Germany

Brice Fabre Institut des politiques publiques (IPP), Paris, France

Sara Flisi European Commission Joint Research Centre, Ispra, Italy

Marzia Freo European Commission Joint Research Centre, Ispra, Italy

Kamel Gadouche Secure Data Access Centre (CASD), Paris, France

Jānis Glazkovs Permanent Representation of Latvia to the European Union, Brussels, Belgium

Ivo Gorissen CBS, The Hague, Netherlands

Dominik Hajduk Government Analytical Unit, Government Office of the Czech Republic, Prague, Czech Republic

Vincent van Hees CBS, The Hague, Netherlands

Emanuel Kulczycki German Centre for Higher Education Research and Science Studies, Berlin, Germany

Roberto Leombruni University of Torino, Torino, Italy

Kamila Lewandowska Science Studies Laboratory, Centre for European Regional and Local Studies, University of Warsaw, Warsaw, Poland

Francisco Lima University of Lisbon, Lisbon, Portugal

Stefano Lombardi VATT Institute for Economic Research, Helsinki, Finland

José Pinto Martins National Statistical Office INE, Lisbon, Portugal

Jaan Masso University of Tartu, Tartu, Estonia

Elena Claudia Meroni European Commission Joint Research Centre, Ispra, Italy

Tatjana Mika German Pension Insurance, Berlin, Germany

Andreea Mitrut University of Gothenburg, Gothenburg, Sweden

Nataša Nater Drvenkar Faculty of Economics and Business, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia

Nikolay Netov University of Sofia, Sofia, Bulgaria

Katja Neugebauer Bank of Portugal, Porto, Portugal

Daphne Nicolitsas Athens University of Economics and Business, Athens, Greece

Alka Obadić Faculty of Economics and Business, University of Zagreb, Zagreb, Croatia

Paolo Paruolo European Commission Joint Research Centre, Ispra, Italy

Joana F. Pimentel Bank of Portugal, Porto, Portugal

Francesco Quatraro University of Torino, Turin, Italy

Sofia Rodrigues National Statistical Office INE, Lisbon, Portugal

Kadri Rootalu Statistics Estonia, Tallinn, Estonia

Pietro Santoleri European Commission Joint Research Centre, Seville, Spain

Paulo Saraiva National Statistical Office INE, Lisbon, Portugal

Ágota Scharle Budapest Institute for Policy Analysis, Budapest, Hungary

Toni Schell Research Data Centre of the Statistical Offices of the Federal States, Duesseldorf, Germany

Alexander Schiersch German Institute for Economic Research (DIW), Berlin, Germany

Maria Schlier University of Copenhagen, Copenhagen, Denmark

Laura Serlenga European Commission Joint Research Centre, Ispra, Italy

Manuel Serrano-Alarcón European Commission Joint Research Centre, Ispra, Italy

Rita Sousa Bank of Portugal, Porto, Portugal

Miroslav Štefánik Institute of Economic Research of the Slovak Academy of Sciences, Bratislava, Slovakia

Nikolaos Theodoropoulos University of Cyprus, Nicosia, Cyprus

Josip Tica Faculty of Economics and Business, University of Zagreb, Zagreb, Croatia

Andreas Tryphonides University of Cyprus, Nicosia, Cyprus

Stefano Verzillo European Commission Joint Research Centre, Ispra, Italy

Koen Vleminckx FPS Social Security, Brussels, Belgium

Matija Vodopivec OECD, Paris, France

Ian Vollbracht European Commission Joint Research Centre, Ispra, Italy

Andres Võrk University of Tartu, Tartu, Estonia

Stefanie Wolter Institute for Employment Research, Nuremberg, Germany

Ewa Zegler-Poleska Science Studies Laboratory, Centre for European Regional and Local Studies, University of Warsaw, Warsaw, Poland

Lennart Ziegler Central European University, Vienna, Austria

João Zilhão National Statistical Office INE, Lisbon, Portugal

Chapter 1
Why This Book?



Fabio Berton and Paolo Paruolo

Abstract This chapter presents the book's motivation, plan and message. The key take-away is that greater systematic access to publicly collected administrative data can generate positive externalities to all stakeholders in the public discourse. The book's main aim is to lower the information asymmetries surrounding this issue.

1 An Executive Summary

This book advocates for improved access to publicly collected microdata. This access can create public value, well beyond the utility of the scientific community. Such data is already being collected at the micro level in order to deliver public service, and its reuse for learning more about what works and for whom can be facilitated without compromising—often actually increasing—data security. The benefits of improved data access for research include increased transparency and trust in the public sector and the optimisation of returns to social investment, among others.

This book's intended audience includes all the stakeholders in the public discourse: the general public, public officials and policymakers, and also our fellow researchers. The book aims to lower asymmetric information about the benefits of access to administrative microdata to learn what works. It focuses on the European Union's (EU) experience, because of the diversity of use-cases in data reuse; however, most of the lessons learned have applicability more generally.

The views expressed in this chapter are those of the authors and do not necessarily reflect the opinion of the institutions of affiliation. The authors thankfully acknowledge useful comments from Ian Vollbracht and Stefano Verzillo on a previous version of the chapter.

F. Berton · P. Paruolo (✉)
European Commission Joint Research Centre, Ispra, Italy
e-mail: paolo.paruolo@ec.europa.eu

F. Berton
e-mail: fabio.berton@ec.europa.eu

© The European Union 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_1

Equality in access to scientific advancement and its benefit is enshrined in Article 27 of the Universal Declaration of Human Rights, United Nations (1948). As reiterated by UNESCO (2021), science, technology and innovation are of vital importance to provide solutions to improve human well-being, advance environmental sustainability and respect for the planet's biological and cultural diversity, foster sustainable social and economic development and promote democracy and peace.

The book is organised into two parts. Part I includes a collection of 27 country chapters (Chaps. 2–28) on the state of play on administrative microdata access in the EU. Part II contains horizontal perspectives on cross-country collaborations (Chap. 29), field perspectives on individual outcomes (Chap. 30), on firms and markets (Chap. 31) and on environmental outcomes (Chap. 32); the final chapter charts the way forward (Chap. 33).

There are ten key take-aways of the book, listed below.

i. Learning what works is important in many different policy fields, ranging inter alia from social investment and public spending, to competitiveness, market regulation, budget (re)allocation and portfolio balancing. It is also relevant to generating evidence about market failures that would require public policies[1] and regarding choices made in everyday life, such as choosing the means of transport from home to work.[2] Chapters 30–32 illustrate some of these benefits.
ii. Field data is a key ingredient whenever evidence about learning what works is based on the modern scientific method.[3] The application of causal methods based on counterfactual impact evaluation techniques (CIE) to administrative microdata sources has great potential: administrative microdata have a large information scope (number of variables), population coverage and time span, they can be linked across multiple data sources (including surveys), and the cost for their reuse is minimal.[4] Moreover administrative microdata lowers the operating costs for any CIE design, including randomised control trials (RCTs). All of this book's chapters uncover various aspects of this.
iii. More evidence on policy effectiveness and on the need for policy intervention arising from an improved access to data is beneficial for the whole of society: (a) the research community would raise the quality and quantity of knowledge and publications; (b) the data-holding institutions would learn the best practices for data safety while providing protected access to data; (c) firms would benefit from greater market transparency; (d) policymakers and citizens would learn

[1] This is sometimes called "problem definition" for policy design.

[2] See McFadden (1974).

[3] Policy-relevant data also contains stakeholders' opinions; here "field data" refers to data generated by individual choices, market clearing, health measurements, etc.

[4] The cost of collection is already covered by their primary use. Causal methods are a fundamental component of empirical sciences, and share their foundations with Probability and Statistics. The field of Econom(etr)ics has lately seen a surge in research contributions on causality (Garg & Fetzer, 2025; van der Klaauw, 2014), driving the so-called credibility revolution in Economics—a movement that led to the awarding of the 2021 Nobel Prize in Economics to some of its initiators, see Hull et al. (2022).

about whether policies work and could make more informed decisions on that basis; and (e) ultimately, trust in public institutions should would be expected to improve.[5]

iv. Many barriers to administrative microdata access still exist despite what has been outlined above, and partly depend on the asymmetric information that exists about its benefits; the country chapters (Chaps. 2–28) offer insights about these obstacles. Hurdles include the perceived increase in the risk of inappropriate use of personal information when granting access to data for research purposes; the misconceptions that alternative solutions to administrative microdata access are less costly and the presumption that Artificial Intelligence (AI) can replace administrative microdata access. Some of these issues are discussed in Chap. 33.

v. Administrative microdata typically encompass the entire population of interest and enable the analysis of distributional effects across various subgroups, such as social categories, geographic locations or income brackets. These effects cannot be investigated with aggregate data, see, e.g., Chap. 32. This allows one not only to measure what works, but also for whom. This can also comprise the starting point of a more structural analysis of the channels that are activated by the policy, thereby also answering the question of why the policy worked, see Heckman (2010).

vi. Policy monitoring can be redesigned as real-time evaluation and, in this way, it can maximise the real-time returns of social investment (policy effectiveness). This generates higher public value than simple monitoring and does so at comparable operating costs. One example of this approach can be found in Chap. 9 on Estonia.

vii. AI is not expected to replace humans in designing, scrutinising and communicating learning what works, but rather to support and to enhance their work.[6] For this reason, human-capital capacity building in the area of policy evaluation will continue to be a critical area for public investment for years to come, see Chap. 32.

The last three take-away messages have a specific application to the EU and to its institutions.

viii. The EU legal framework, associated with the General Data Protection Regulation (GDPR, EU (2016)), is often perceived as a limitation to possible administrative microdata access. The use-cases that are collected in the country chapters (Chaps. 2–28) prove that these obstacles are not insurmountable and that there exist an array of options and good practices from among which Member States can choose.

ix. Policy interventions always generate administrative microdata (implementation data); their collection and preparation for research and reporting purposes

[5] (a) See Chap. 29; (b) see Gong et al. (2022) on the use of differential privacy techniques for US census data; (c) see Chap. 32; (d) see, e.g., Chap. 9; (e) see Carattini et al. (2024).

[6] See Paruolo (2005).

can largely be automated, thereby generating efficiency and cost gains when compared to ex-post surveys. Survey data can also be linked to the administrative microdata to provide additional information not present there. The systematic use of administrative microdata hence appears in line with process simplification while enabling the use of the best available data, see the European Commission's Communication on Implementation and Simplification (EU, 2025) and its Better Regulation Agenda (EU, 2015); (EU, 2021a).

x. Some EU Member States have devoted considerable effort to easing administrative microdata reuse (see Chaps. 2–28). A persisting lack of coordination among EU Member States, and of policies that support a more systematic access to administrative microdata, may induce the abandonment of good practices. Hence, public policies in support of administrative microdata reuse and coordination appear justified, see Bijlsma et al. (2021).

The remainder of this chapter is organised as follows: Sect. 2 discusses the book's motivation. Section 3 explains what the book is (not) about, while Sect. 4 offers the rationale for the choice of material. Section 5 lists some common misconceptions about administrative microdata and policy evaluation that may hinder their more widespread use in supporting policymaking. Finally, Sect. 6 describes the ways in which the book can be read and the different types of readership envisioned.

2 Motivation

2.1 Evidence for Policy

Learning what works, for whom, and–where possible–why, is essential to enhancing the public value generated by policy interventions, see Besharov et al. (2017). The following examples help to illustrate the use of causal evidence for policy.

Causal evidence can help to target policy. For instance, in a study on COVID-19, Chetty et al. (2023) report evidence that the pandemic reduced spending among high-income individuals, lowered revenues for SMEs in dense urban areas and caused job losses among low-wage earners. This evidence is informative for the design of post-pandemic measures.

Another example comes from the evidence in Pinotti (2017), who found that the legalisation of migrants in Italy more than halved their crime rate; this evidence can inform future policy. In another causal study, Berton et al. (2018) found that a 10% reduction in bank credit to Italian enterprises led to a 3.6% decline in employment between 2008 and 2012; this provided rationale for measures such as the quantitative easing.[7]

Causal evidence can also uncover the relation between the timing of policy and their effectiveness. For instance, Gillingham and Stock (2018) review evidence about

[7] The Federal Reserve started this policy in 2008 and the European Central Bank in 2015.

policies aimed at reducing greenhouse gas emissions; some of them may initially appear inefficient, given their high upfront costs (a concern that is particularly salient for countries with tight public budgets), but their dynamic long-run effects can be substantial, ultimately making these interventions cost-effective.

Another example is the evidence about the training programmes provided by Card et al. (2018), who showed that training increases the probability of employment two to three years after completion; judging trainings only by their short-term outcomes would, hence, risk to discard effective training courses. Finally, comparing causal policy effects in different countries offers an opportunity to discuss the differential roles of markets and institutions, see Bertheau et al. (2023).

As these examples show, knowing what works provides not only accountability regarding policy design, but also creates public value. This value is widely recognised and emphasised in official documents issued by both national and international institutions. For instance, the European Commission's (EC) Better Regulation Agenda—a package of measures designed to enhance the quality and effectiveness of EU laws and policies—aims to ensure that EU initiatives are "based on the best available evidence and analysis" (EU, 2015, p. 7).

The accompanying Better Regulation Guidelines further underline that "evaluations support decision-making and contribute to strategic planning and the design of future interventions" (EU, 2021a, p. 24). Similarly, the report on evaluation activities within the EC notes that "Evaluation has become a key tool to steer EU policymaking" (Anglmayer, 2020, p. 8).

In line with the EC's approach, the Organisation for Economic Co-operation and Development (OECD) also underscores the centrality of evidence-based policymaking, stating that "policy evaluation contributes fundamentally to sound public governance. It can help governments improve the design and implementation of public policies that can, in turn, lead to prosperity for their country and well-being for its citizens" (OECD, 2020b, Executive Summary).

Similarly, the World Bank highlights the importance of evaluation for effective policy design through its Competitiveness Policy Evaluation Lab, which promotes the use of evaluation to maximise the impact of key projects in policy areas such as firm performance and productivity, entrepreneurship, innovation and SME finance.[8] Finally, the International Labour Organization (ILO) also places a strong emphasis on policy evaluation; the ILO operates its own evaluation office, and is committed to both fostering and disseminating a culture of evaluation (ILO, 2023).[9]

[8] See: https://www.worldbank.org/en/programs/competitiveness-policy-impact-evaluation-lab. See also the World Bank's recent report on the usefulness of making public data available for research, Hasnain et al. (2024).

[9] See: https://www.ilo.org/about-ilo/how-ilo-works/accountability-and-transparency/evaluation-office.

2.2 Evidence Generation and Data

It is simple to observe that no applied scientific evidence is available without data, given the role of observation in the modern scientific method (Crato & Paruolo, 2019). This also holds when one takes a broader definition of evidence for policy as in OECD (2020a), one that also encompasses knowledge stemming from types of policy evaluation that are different from CIE, given the role played by data in these processes as well.

The EC emphasises the importance of data in informing policy decisions stating that "there is a strong potential for objectives of general interest in the use of data" including "the provision of public services, or public policy making" (EU, 2022, recital 45), and adding that "data is the lifeblood of economic development (...) enabling better policy making and upgrading government services" (EU, 2020, p. 2).

This is echoed by the OECD who states that "governments are also recognizing that the data they hold is a strategic asset that can be used to generate evidence to inform the performance of policies" (OECD, 2020a, p. 15), and in particular that "administrative data permit deriving detailed and high-quality policy insights that can be used to improve policy design and implementation" (OECD/EC, 2025, p. 9).

Similarly the World Health Organization (WHO) asserts that "for 71 years, the WHO has had one vision: the highest attainable standard of health for all people in all countries. Reliable data are vital for assessing whether we are making progress towards that vision" (World Health Organization, 2019, p. v).

Ideal data sources to support policymaking with sound empirical evidence are not always available; in some cases, they do not even exist. More frequently, such data exist but they are not easily accessible. The EC has been actively promoting the availability of data to the research community, recognising its potential to drive innovation, economic growth and social progress.

One of the key initiatives in this regard is the General Data Protection Regulation (GDPR), which came into effect in 2018 (EU, 2016). While the GDPR is often seen only as a regulatory framework aimed at protecting personal data, its Article 89 recognises a privileged access to data for the purposes of research, based on the superior and collective interest in knowledge and subject to the implementation of appropriate safeguards to protect the rights and freedoms of data subjects. This means that researchers in the EU have legally recognised access to a wide range of data, including personal microdata. These data can be used to conduct studies on topics such as, but not limited to, health, economics and other social sciences.

The EC has also launched the EU Data Strategy (EU, 2020) that aims to create a single EU data space. Together with the EU Data Act (first introduced with Directive 2020/1828 and amended thereafter through Regulation 2023/2854 (EU, 2023) and the Data Governance Act (EU, 2022)), the EU Data Strategy recognises the importance of data in driving innovation and economic growth and sets out a framework for the creation of a data-driven economy.

The EU Data Strategy includes measures such as the creation of data sharing platforms, the development of data standards and the establishment of a framework

for data governance. The EU Data Strategy also emphasises the importance of making data available to researchers, including through the creation of data repositories and the provision of funding for data-driven research projects.

A key initiative within the EU Data Strategy concerns data spaces, which refer to the creation of shared data infrastructures that allow for the secure and efficient sharing of data between different stakeholders. The data spaces initiative aims to create a federation of decentralised data ecosystems that can be used by researchers, businesses and public organisations to access and share data. Data spaces in the making include, but are not limited to, health, the green deal, agriculture, finance, skills and public procurement.

Furthermore, the Open Data Directive (EU, 2019b) aims to make public sector data available for reuse. The directive requires EU Member States to make public sector data available in a machine-readable format, free of charge and with minimal restrictions on reuse. This has led to the creation of a wide range of national open data portals—as this book documents—providing access to data on topics such as transportation, education and health care.

The directive on Infrastructure for Spatial Information in the EU (INSPIRE, see EU (2007)) has also been instrumental in making large amounts of spatial data available to researchers, which has enabled them to conduct studies on topics such as climate change, urban planning and environmental sustainability.

In addition to these initiatives, the EC also indirectly promotes the access to data through several funding programmes that are aimed at supporting data-driven research. The most important programme is the Horizon 2020 programme (EU, 2013), afterwards renewed as Horizon Europe (EU, 2021b), and this has provided funding for research projects that use data to address societal challenges such as climate change, health care and transportation.

2.3 Choice of Material

In spite of the EC's efforts to broaden and ease the availability of data, researchers often face difficulties in administrative microdata access in practice; this is where the present book aims to contribute by lowering information asymmetries.

The book makes four specific choices in order to pursue that aim. First, the book has an EU focus, even though most ideas and solutions have general validity. The EU is an interesting case study; in fact, it has a system of shared rules (like the GDPR) and institutions (like the EC), with different national interpretations and applications; this represents an ideal environment in which to compare different solutions and practices. Moreover, the EU also represents a natural laboratory in which different policy interventions, or variations of a policy, are applied to common problems.

Second, the book focuses on administrative microdata, i.e. "data collected by public entities for their service activities" (Crato & Paruolo, 2019, p. v) and tracked at the most elementary statistical unit (individuals, households, firms, etc.); again, most of the ideas and solutions also apply to similar microdata in the private sector.

Section 4 below explains why administrative microdata are best-placed to feed causal analysis, in particular in association with CIE methods.

Third, this book is intended for a diverse range of readers, including policymakers, who share an interest in public utility and social value. The book takes the data-users' perspective, and the researchers' one in particular. In fact, researchers play an instrumental role in transforming data into knowledge in the process of generating public value; hence, this specific perspective is a natural one. However, the book is written in non-technical language, so as to be understandable by the general public.

Fourth, reliable evidence depends on reliable data, researchers and methods.[10] The present book does not discuss methods in detail, but rather builds on the premise that CIE techniques can generate policy-relevant evidence and knowledge,[11] and they deliver causal evidence when applied properly. This set of methods has implications in terms of which data are most suitable for providing sound evidence for policymaking.

3 Scope of the Book

Part I of the book explains how to get access to microdata from administrative sources in the 27 EU Member States for research purposes. It describes the formal procedures, the offices a researcher needs to get in touch with, the criteria that a data access request is assessed with, the access methods (whether from remote or on-site, via a hardware key or multiple authentication, etc.), the monetary costs and the—often unknown to people from outside each single data-holding institution—practical tips to speed up the process of data access.

In spite of the common legislative framework, there are 27 different combinations of options for administrative microdata access, one per EU Member State. The book does not attempt to rank them according to any given criterion; instead, it aims to document current practices of administrative microdata access and safeguards for data protection. Any data-holding institution can derive lessons learned from this, in a context of continuously evolving technology and knowledge needs.[12]

[10] See also the "five safes" cited in Chap. 29.

[11] Introductory material on CIE can be found in Cunningham (2021), EU (2019a) and Gertler et al. (2016); see Hull et al. (2022) for a historical perspective. Complete presentations of CIE can be found in, e.g. Angrist and Pischke (2008, 2014) and Imbens and Rubin (2015).

[12] The Open, Useful and Reusable data (OURdata) index by the OECD assesses open data policies at the national level. It is an unweighted composite indicator of nine subpillars in the areas of data availability, data accessibility and government support to data reuse. As such, it can be compared across countries. However, open data "refer[s] to non-discriminatory data access and sharing arrangements, where data is machine readable and can be accessed and shared, free of charge, and used by anyone for any purpose subject, at most, to requirements that preserve integrity, provenance, attribution, and openness" (OECD, 2021). The OURdata initiative therefore refers to aggregate or fully anonymised microdata not specifically conceived for research, thereby excluding most of the data sources with which this book is concerned. A comparable reasoning applies to the Open Data Maturity Report by the EU (2024).

As anticipated, the book takes the data-user perspective. To do so, the editors have invited a total of approximately 70 contributors to describe how exactly to file a request for administrative microdata access in the Member State that they know best and/or to provide examples of their use. They are experts of the use of administrative microdata from universities, statistical offices, central banks, national and international organisations or research centres. The result is an unprecedented wealth of formal and practical information that, hopefully, can help and inspire researchers and data-users at large.

This book does not aim to be exhaustive. Only a minority of EU Member States have developed a centralised and standardised data access procedure for all of the administrative sources available at the micro level. In the other Member States, access is more or less fragmented, with procedures that are unique for a single data-holding institution. In these cases, for reasons of space, the country chapters typically proceed by providing examples concerning specific domains. Irrespective of this aspect, all of the 27 country chapters identify directions for further improvement.

Beyond data access procedures, Part II of the book aims to document the relevance of administrative microdata in several policy fields, including labour, health, education, agriculture, productivity and innovation. Research originating from administrative microdata is shown to be of crucial importance in all policy fields.

One complication for evaluations is that EU policies are superimposed to 27 different national policies. Policy-relevant research should therefore study more EU Member States together to learn from different national institutional settings, see Bertheau et al. (2023). There is no EU standard on administrative microdata collection allowing for cross-border data merging, and data wrangling is thus even more difficult for cross-country projects. In this area, Chap. 29 describes international data sharing, cross-country data linkages and fora for exchange. The description of such international experiences makes clear that interactions (among people, between people and data) emerge as a crucial component of the development of policy-relevant knowledge.

Despite the richness of the use-cases included, the book does not cover a number of areas. Firstly, the technical legal aspects of data access and protection are beyond the book's scope. Secondly, although replicability is intrinsic to science, data cheating, (deep) fakes, etc., are only marginally discussed in this book. Third, the book does not document use-cases beyond the EU. Lastly, while administrative microdata are collected at different government levels, including local ones, the book's focus is primarily on the EU and national levels. All these limitations are dictated by space and time constraints.

4 Why Administrative Data?

Learning what works and for whom requires scientific evidence, which is the result of appropriate methods being applied to reliable data. This section lists favourable features of administrative microdata for causal CIE methods.[13]

Causality has a central role in science for policy, as clarified by the following example from medicine. Coffee drinking is positively associated with lung cancer; once broken down between smokers and non-smokers, this association is found to hold only among smokers (e.g. see the meta-analysis by Galarraga and Boffetta (2016)). A reason for this association can be that smokers usually also drink coffee, and smoking is a cause of lung cancer; in this case any policy aiming to reduce coffee consumption would not have any effect on lung cancer.

This highlights the need to base policy decisions (and hence policy support) on causal evidence whenever possible. This is all the more important at the present time when policies need to adapt to a fast-changing environment and ambitious targets.

Scientific evidence needs to be based on the best available data, in terms of relevance, accuracy and coverage.[14] In this context, the best data is usually obtained by merging administrative data—covering for instance all characteristics contained in registries, such as age, working history and education of a human population—with survey data on perceptions, which are not contained in administrative data.

Administrative microdata have a higher reliability potential compared to other sources of data, and survey data in particular. The monetary value of administrative microdata is already well understood by the private sector.[15] Moreover, administrative microdata are of great value for causal CIE analysis, and therefore for policy-relevant support; the following sections list six reasons for this: coverage, bias, costs, linkage, timeliness and time span.

4.1 Coverage and Bias

Administrative microdata have a large coverage because they are collected by public authorities for delivering service. Administrative microdata often cover the entire reference population, while surveys—for cost-related reasons—are administered to samples from the population. A high coverage is desirable for two reasons: it allows

[13] Garfinkel (2018) discusses advantages and challenges of working with "organic data" (or "found data"), which include administrative microdata and other non-classical data sources, rather than with "designated data" (or "made data") from traditional sources like surveys.

[14] See Crato and Paruolo (2019) for a relation to big data, characterised by volume, variety and velocity.

[15] Balance sheets of firms are public administrative microdata that is collected and sold by the private sector. Targeted advertisement on Internet platforms is based on microdata provided by users, which however does not come from administrative microdata sources.

for subpopulation analysis—crucial to know for whom the policy is working or not—and it offers a more precise quantification of the computed causal effect, a necessary premise for cost-benefit analysis.[16]

Administrative microdata are not subject to the recall or reporting bias that is experienced in surveys. The administrative microdata's elementary information is collected in real time and does not originate from a questionnaire in which the respondents might find it difficult to remember what happened years prior, or may even prefer to hide pieces of information that they deem sensitive (like earnings).

Recent advances use linked experimental data with administrative microdata to uncover the long-term effects of policies, see Athey et al. (2019, 2025).

4.2 Cost and Linkage

The cost of administrative microdata collection is covered by their primary use (unlike for nation-wide surveys).

The elementary statistical unit in administrative microdata is almost always associated with a unique identifier—e.g. the social security number or the VAT number—shared across different archives. This implies that administrative sources can be linked to each other—and to survey data where they are also collected using the same identifiers—thus providing data-users with a larger information content and knowledge potential.[17]

4.3 Timeliness and Time Span

Administrative microdata are collected continuously, hence their availability for policy-relevant questions is close to real time, provided that the necessary data checks are conducted properly.

Each single statistical unit (an individual, a household, a firm) can be followed easily over time with administrative microdata, including over long time horizons, like decades or generations as, e.g. in Abrahamsson et al. (2025).

Administrative microdata are, therefore, in a position to satisfy the FAIR principles first introduced in Wilkinson et al. (2016) and recalled by the European Strategy for Data (EU, 2020), according to which data should be Findable, Accessible, Interoperable and Reusable.

[16] The entire population can be considered as a sample from a superpopulation, or can be treated as a finite population, see, e.g., Abadie et al. (2023).

[17] This also applies when using multiple data sources, provided each pair of datasets shares one unique identifier, see Crato and Paruolo (2019).

5 Common Misconceptions

Despite all of administrative microdata's positive characteristics, barriers still exist that prevent their full use. There are several reasons for why this is: one reason is the lack of information about either associated benefits or about the cost of inaction; another reason is that there are some common misconceptions about administrative microdata and CIE methods.

5.1 Relative Costs

Does administrative microdata cost more than data from a dedicated survey? No: dedicated surveys have high monetary costs and need to be repeated for every round of data collection, while administrative microdata already exist. While surveys generate "research-ready" data (this is an advantage), administrative microdata need to be cleaned and prepared for research. However, the data-preparation investment for administrative microdata has decreasing costs over time; the associated computer codes can typically be run nearly unchanged at every new update of the raw data, making the marginal costs eventually very low.

5.2 Turnaround Times

One common misconception is that policymaking requires fast evidence that CIE based on administrative microdata cannot provide. In fact both CIE and simple monitoring often boil down to computer codes applied to the same microdata; this implies that the computational time required by CIE analyses, per se, is not longer than the one for monitoring indicators.[18]

Preparing administrative microdata for research requires initial investments, however. This can be largely automated for nearly real-time flows of new data. Conversely, dedicated surveys have high costs in terms of the design of the questionnaire, its validation and administration. Thus, contrary to commonly held beliefs, CIE on administrative microdata can be faster and cheaper than CIE on a dedicated survey. Similarly, CIE on administrative microdata can have the same turnoround time as monitoring indicators, while also generating a larger public value.

[18] The majority of CIE methods have the same computational time order of magnitude as the aggregation of indicators.

5.3 Data Protection Issues

Administrative microdata access raises data protection concerns that surveys can address through informed consent, which involves seeking interviewees' permission to use their data for additional research. This possibility offers surveys both benefits and drawbacks, including increased time to complete questionnaires and a potential reduction in the number of re-usable observations in the sample.[19]

The interplay between personal and societal utility of administrative microdata reuse is debated in economic theory, see, e.g., Acquisti et al. (2016) and Jones and Tonetti (2020). On the side of formal privacy techniques (Garfinkel, 2018), continuous improvements in differential privacy (Dwork et al., 2014) and other techniques (Hurtado-Ramírez et al., 2023) are advancing the possibility to protect data privacy while also giving access to that data.[20]

From a practical perspective, the US report "The Promise of Evidence-Based Policymaking", US (2017), found that granting access to administrative microdata for research often involves stronger, rather than weaker, safeguards for personal data. This is because implementing privacy-enhancing measures represents a significant improvement over the alternative of no access, which can leave data unprotected. The country-specific chapters (Chaps. 2–28) are in line with this observation.

6 Readers' Guide

The book lends itself to a three-level reading; this section describes possible ways in which to read this book.

First, and most straightforwardly, the book provides a user's guide on how to get access to data in each of the 27 EU Member States. Chapters 2–28 cover the main administrative microdata sources and provide a starting point in the search for data.

Second, the set of country chapters as a whole (Chaps. 2–28) represents an inventory of access methods where the national data-holding institutions, who are seeking to improve the access to the data for which they are responsible, can draw and learn from. All of these administrative microdata access methods fulfil the requirements of the European legislative framework, and of the GDPR in particular.

Third, the policymakers at any government level can learn about the relevance of research from administrative microdata sources in Chaps. 29–33. They can identify specific areas of policy intervention that aim towards improved access to administrative microdata.

---

[19] The interviewee's choice not to accept data reuse introduces a further selection and, hence, also changes the reference population.

[20] For instance, the US Census has adopted differential privacy techniques for public products from the 2020 Census, see Gong et al. (2022). The field is in rapid evolution and challenges are still being resolved, see Drechsler (2023).

The three-level reading implies diverse types of readers. Researchers, and data-users in general, can benefit from the book for guidance on where to get access to data. Officers and representatives from the data-holding institutions can learn about the best practices in safe data access. Policymakers can learn about the value of data and of causal evidence.

However, the book's intended and primary readership is all citizens; they are the ultimate beneficiaries of policies and the book intends to inform them about how robust causal evidence can be produced. Empirical evidence shows that they are sensitive to policies that are based on causal evidence (Carattini et al., 2024), and this can enhance their trust in public institutions, and ultimately in democracy. This is ultimately why this book project was conceived.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics, 138*(1), 1–35. https://doi.org/10.1093/qje/qjac038

Abrahamsson, S. S., Bütikofer, A., Løken, K. V., & Page, M. E. (2025). *Sources of generational persistence in the effects of early-life health interventions* (Working Paper, 33612). National Bureau of Economic Research. https://doi.org/10.3386/w33612

Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature, 54*(2), 442–92. https://doi.org/10.1257/jel.54.2.442

Anglmayer, I. (2020). *Evaluation in the European Commission. Rolling check-list and state of play* (Technical report). European Parliamentary Research Service. Retrieved February 1, 2026, from https://www.europarl.europa.eu/RegData/etudes/STUD/2020/654170/EPRS_STU(2020)654170_EN.pdf

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering metrics: The path from cause to effect*. Princeton University Press.

Athey, S., Chetty, R., & Imbens, G. W. (2025). *The experimental selection correction estimator: Using experiments to remove biases in observational estimate.* (Working Paper 33817). National Bureau of Economic Research.

Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2019). *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely* (Working Paper, 26463). National Bureau of Economic Research.

Bertheau, A., Acabbi, E. M., Barceló, C., Gulyas, A., Lombardi, S., & Saggio, R. (2023). The unequal consequences of job loss across countries. *American Economic Review: Insights, 5*(3), 393–408.

Berton, F., Mocetti, S., Presbitero, A., & Richiardi, M. (2018). Banks, firms and jobs. *The Review of Financial Studies, 31*, 2113–2156. https://doi.org/10.1093/rfs/hhy003

Besharov, D. J., Baehler, K. J., & Klerman, J. A. (2017). *Improving public services: International experiences in using evaluation tools to measure program performance.* Oxford University Press.

Bijlsma, M., van der Klaauw, B., & Knoef, M. (2021). The government's data agenda should also focus on the data itself. *ESB, 106*, 388–391. Retrieved February 1, 2026, from https://esb.nu/de-data-agenda-van-de-overheid-dient-zich-ook-op-de-data-zelf-te-richten/ (in Dutch)

Carattini, S., Dur, R., & List, J. (2024). Policy evaluation and the causal analysis of public support. *Science, 386*, 490–492. https://doi.org/10.1126/science.adp7497

Card, D., Kluve, J., & Weber, A. (2018). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association, 16*, 894–931. https://doi.org/10.1093/jeea/jvx028

Chetty, R., Friedman, J. N., Stepner, M., & Opportunity Insights Team. (2023). The economic impacts of Covid-19: Evidence from a new public database built using private sector data. *The Quarterly Journal of Economics, 139*(2), 829–889. https://doi.org/10.1093/qje/qjad048

Crato, N., & Paruolo, P. (2019). *Data-driven policy impact evaluation: How access to microdata is transforming policy design.* Springer. https://link.springer.com/book/10.1007/978-3-319-78461-8

Cunningham, S. (2021). *Causal inference: The mixtape* (1st ed.). Yale University Press. Retrieved February 1, 2026, from https://mixtape.scunning.com

Drechsler, J. (2023). Differential privacy for government agencies—Are we there yet? *Journal of the American Statistical Association, 118*(541), 761–773. https://doi.org/10.1080/01621459.2022.2161385

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science, 9*(3–4), 211–407. https://doi.org/10.1561/0400000042

EU. (2007). *Directive on establishing an infrastructure for spatial information in the European community (inspire)* (Directive 2007/2/EC). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02007L0002-20241126

EU. (2013). *Regulation establishing Horizon 2020—The framework programme for research and innovation* (Regulation (EU) No 1291/2013). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013R1291

EU. (2015). *Better regulation for better results—An EU agenda* (COM(2015) 215 final). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52015DC0215

EU. (2016). *General data protection regulation* (Regulation (EU) 2016/679). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

EU. (2019a). *Advanced counterfactual evaluation methods—Guidance document*. https://op.europa.eu/en/publication-detail/-/publication/11968bbb-fac9-11e9-8c1f-01aa75ed71a1/language-en

EU. (2019b). *Open data and the reuse of public-sector information* (Directive (EU) 2019/1024). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L1024

EU. (2020). *A European strategy for data* (COM(2020) 66 final). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066

EU. (2021a). *Better regulation guidelines* (SWD(2021) 305 final). https://commission.europa.eu/document/download/d0bbd77f-bee5-4ee5-b5c4-6110c7605476_en?filename=swd2021_305_en.pdf

EU. (2021b). *Regulation establishing Horizon Europe—The framework programme for research and innovation* (Regulation (EU) 2021/695). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32021R0695

EU. (2022). *European data governance act* (Regulation (EU) 2022/868). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R0868

EU. (2023). *Data act* (Regulation (EU) 2023/2854). https://eur-lex.europa.eu/eli/reg/2023/2854

EU. (2024). *Open data maturity report 2024*. https://doi.org/10.2830/6995777

EU. (2025). *A simpler and faster Europe: Communication on implementation and simplification* (COM/2025/47 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52025DC0047

Galarraga, V., & Boffetta, P. (2016). Coffee drinking and risk of lung cancer—A meta-analysis. *Cancer Epidemiology, Biomarkers and Prevention, 25*(6), 951–957. https://doi.org/10.1158/1055-9965.EPI-15-0727

Garfinkel, S. L. (2018). Privacy and security concerns when social scientists work with administrative and operational data. *The Annals of the American Academy of Political and Social Science, 675*(1), 83–101. https://doi.org/10.1177/0002716217737267

Garg, P., & Fetzer, T. (2025). *Causal claims in economics* (Technical report). https://doi.org/10.48550/arXiv.2501.06873

Gertler, P., Martinez, S., Premand, P., Rawlings, L., & Vermeersch, C. (2016). *Impact evaluation in practice*. Inter-American Development Bank and World Bank. https://doi.org/10.1596/978-1-4648-0779-4

Gillingham, K., & Stock, J. (2018). The cost of reducing greenhouse gas emissions. *Journal of Economic Perspectives, 32*, 53–72. https://doi.org/10.1257/jep.32.4.53

Gong, R., Groshen, E. L., & Vadhan, S. (2022). *Differential privacy for the 2020 census: How can we make data both private and useful*? Retrieved February 1, 2026, from https://privacytools.seas.harvard.edu/blog/differential-privacy-2020-census-how-can-we-make-data-both-private-and-useful

Hasnain, Z., Khurshid, A., Lundy, T., & Rogger, D. (2024). *Government analytics in Europe: Making public data count*. The World Bank Group. Retrieved February 1, 2026, from https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099082924113516503

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic literature, 48*(2), 356–398. https://doi.org/10.1257/jel.48.2.356

Hull, P., Kolesár, M., & Walters, C. (2022). Labor by design: Contributions of David Card, Joshua Angrist, and Guido Imbens. *The Scandinavian Journal of Economics, 124*(3), 603–645. https://doi.org/10.1111/sjoe.12505

Hurtado-Ramírez, D., et al. (2023). *Technological enablers for privacy preserving data sharing and analysis* (Technical report). JRC. https://doi.org/10.2760/427718

ILO. (2023). *ILO results-based evaluation strategy 2023–2025*. Retrieved February 1, 2026, from https://www.ilo.org/publications/ilo-results-based-evaluation-strategy-2023-25

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the Economics of data. *American Economic Review, 110*(9), 2819–58. https://doi.org/10.1257/aer.20191330

McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics, 3*, 303–328. https://doi.org/10.1016/0047-2727(74)90003-6

OECD. (2020a). *Building capacity for evidence-informed policy-making. Lessons from country experiences*. https://doi.org/10.1787/86331250-en

OECD. (2020b). *Improving governance with policy evaluation: Lessons from country experiences*. https://doi.org/10.1787/89b1577d-en

OECD. (2021). *Recommendation of the Council on enhancing access to and sharing of data* (OECD/LEGAL/0463). Retrieved February 1, 2026, from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0463

OECD/EC. (2025). *Counterfactual impact evaluations of active labour market policies: Lessons from using linked administrative data*. https://doi.org/10.1787/bea7860e-en

Paruolo, P. (2005). Automated inference and the future of econometrics: A comment. *Econometric Theory, 21*(1), 78–84. https://doi.org/10.1017/S0266466605050061

Pinotti, P. (2017). Clicking on heaven's door: The effect of immigrant legalization on crime. *American Economic Review, 107*, 138–168. https://doi.org/10.1257/aer.20150355

US. (2017). *The promise of evidence-based policymaking: Report of the commission on evidence-based policymaking*. United States Commission on Evidence-Based Policymaking. Retrieved February 1, 2026, from https://digital.library.unt.edu/ark:/67531/metadc1228332/

UNESCO. (2021). *UNESCO recommendation on open science*. United Nations Educational, Scientific and Cultural Organization. https://doi.org/10.54677/MNMH8546

United Nations. (1948). *Universal declaration of human rights*. United Nations. Retrieved February 1, 2026, from https://www.un.org/en/about-us/universal-declaration-of-human-rights

van der Klaauw, B. (2014). From micro data to causality: Forty years of empirical labor economics. *Labour Economics, 30*, 88–97. https://doi.org/10.1016/j.labeco.2014.06.009

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data, 3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

World Health Organization. (2019). *World health statistics 2019*. Retrieved February 1, 2026, from https://iris.who.int/bitstream/handle/10665/324835/9789241565707-eng.pdf

Fabio Berton is Policy Analyst on impact evaluation at the Competence Centre for Microeconomic Evaluation at the Joint Research Centre of the European Commission. He is also Associate Professor of Economic Policy at the University of Torino, and Research Fellow at CIRET, IZA and LABORatorio R. Revelli. His research focuses on applied labour economics and on the evaluation of labour policies in particular. Fabio has publications on *The Economic Journal*, *The British Journal of Industrial Relations*, *The Oxford Bulletin of Economics and Statistics*, *The Review of Financial Studies* and *The Journal of Economic Inequality* among the others.

Paolo Paruolo is an econometrician, currently Head of the European Commission Competence Centre on Microeconomic Evaluation since 2016. He holds a Ph.D. in Econometrics from the University of Copenhagen, and he has been Full Professor of Econometrics at the University of Insubria, Varese, Italy, and previously at the University of Bologna. His research focuses on econometric theory, time series analysis and counterfactual methods. He has published in *Econometrica*, *Journal of Econometrics*, *Econometrics Journal, Econometric Theory*, *Journal of the Royal Statistical Society Series A*, *Oxford Bulletin of Economics and Statistics*, among others. He has written and edited several books, including "Data-driven policy impact evaluation: How access to microdata is transforming policy design", Crato, Paruolo (eds 2019).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part I
EU Experiences

Chapter 2
Working with Large Administrative Datasets from Austria



Omar Bamieh and Lennart Ziegler

Abstract This chapter discusses opportunities that are available for researchers to use linked microdata from Austria. While access to register data has been limited historically, two recent initiatives have substantially expanded the scope for empirical research: the Austrian Microdata Center (AMDC) and the Austrian Labor Market Database (*Arbeitsmarktdatenbank*, AMDB). In this chapter, we outline the available datasets and describe how researchers might both access and work with these sources. One key novelty presented by the initiatives is that linkages across registers from different sources are now feasible, thereby opening up new avenues for interdisciplinary research. Having created the infrastructure for microdata access, future efforts should focus on extending the data catalogue to cover more fields of research and data from previous years.

1 Introduction

Austria provides a number of high-quality and large-scale administrative datasets; however, access for research purposes is often restricted by the data-hosting ministries and by government agencies.[1] Public institutions have the obligations to facilitate (academic) research, where possible, but decisions about access are often made at the discretion of the respective institution. This limits the extent to which independent research can be conducted and raises uncertainty about data availability. Realised

[1] One example is the Austrian Social Security Administration, which provides a list of recent cooperations on their website (https://www.sozialversicherung.at/cdscontent/?contentid=10007.850727&portal=svportal)

O. Bamieh (✉)
Andrés Bello University, Santiago, Chile
e-mail: omar.bamieh@unab.cl

L. Ziegler
Central European University, Vienna, Austria
e-mail: ZieglerL@ceu.edu

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_2

research projects are typically commissioned by policymakers, even though individual researchers can suggest project ideas and can apply for data access. Another difficulty is that universal person identifiers do not exist in Austria, unlike in countries like Denmark. To that end, a separate government agency needs to assign encrypted personal identifiers based on confidential personal data (name, birthdate, address, social security number)[2] in order to merge data across different organisations. In this chapter, we discuss two recent data initiatives that aim to overcome these limitations and extend microdata access for individual research purposes: the Austrian Microdata Center (AMDC) and the Austrian Labor Market Database (*Arbeitsmarktdatenbank*, AMDB). Sect. 2 provides an overview of the Austrian Microdata Center (AMDC), which is hosted by Statistics Austria and offers a secure work environment for researchers to access and link microdata from various sources. The centre also allows for the merging of data provided externally, alongside an extensive list of readily available datasets. In Sect. 3, we outline the Austrian Labor Market Database (AMDB), which combines spell-level data from the Social Security Administration and the Public Employment Service in Austria.[3] When used in conjunction, the matched employer-employee records offer a comprehensive coverage of the Austrian labour market.

2 Austrian Microdata Center (AMDC)

Statistics Austria established the Austrian Microdata Center (AMDC) in 2022 in order to provide a secure work environment for researchers to access and link microdata from various sources. Accredited research institutions (even though outside of Austria) can submit applications for research projects. The researcher gains remote data access against a usage fee when a proposed project is deemed feasible, and provided it fulfils all data protection requirements.[4] There is no requirement to carry out the analysis within the premises of Statistics Austria.[5] All projects that fulfil the requirements are approved; there is no ranking based on scientific merit. The AMDC data catalogue contains a wide range of microdata that is continuously being updated with more recent observations and additional datasets. All data are well documented in German. Moreover, researchers can supply external data and this can be linked

[2] The government agency provides additional information on the procedure (in German): https://www.bmf.gv.at/ministerium/aufgaben-und-organisation/Stammzahlenregisterbehoerde/Bereichsspezifische-Personenkennzeichen-(bPK)/Verschl%C3%BCsselte-bereichspezifische-Personenkennzeichen-(vbPK).html

[3] By spell-level data, we mean data about a duration or period of work, time spent out of the labour market etc.

[4] Total usage costs depend on the specific project requirements. See the AMDC cost catalogue for more information: https://www.statistik.at/fileadmin/pages/1805/Katalog_der_Serviceleistungen.pdf

[5] The specific requirements for accreditation and research proposals are defined in the Federal Statistics Act (BStatG § 31).

on the person- or firm-level to other datasets within the work environment if the appropriate identifiers have been provided.

There are two broad categories of data available in the data catalogue.[6] First, Statistics Austria has a legal obligation to collect data on a wide range of topics. Many surveys are carried out by Statistics Austria at different time intervals and are often available through the AMDC.[7] For example, the Microcensus (and adjunct modules) are conducted on a regular basis and this provides large-scale, representative survey data to researchers. Second, various government agencies report register data to Statistics Austria, thereby enabling the office to both compile and to analyse federal statistics. This means that certain types of information are available for every citizen and for every firm in Austria. These registers allow research in various fields, including health, education, family relations, employment, income and housing. Most of these datasets are available for recent years, while some also include records dating back to the early 2000s. Data from previous years are rarely accessible, due to the lack of previously undertaken data collection efforts and the absence of appropriate identifiers. Data are normally updated once the most recent records become available, which can take up to three years. It is possible to link all registers on the person- and/or firm-level. In what follows, we will outline some of the key administrative datasets in the AMDC and will discuss their potential for future uses in research.

2.1 Overview of Datasets

The **Employment Statistics** (*Abgestimmte Erwerbsstatistik und Registerzählung*) combines general population characteristics with employment records, both on the worker- and firm-level by providing annual information about every person residing in Austria. The data have been collected every year at the end of October since 2009. Individual demographic characteristics include gender, age, nationality and marital status. The dataset provides a detailed classification of the person's labour market status, and, if applicable, information about their employer's sector and firm size. Person and firm identifiers allow researchers to track the labour market flows of workers across employers, a feature that has been used extensively in labour market research, in order to study wage determinants for instance (Abowd et al., 1999). Municipality identifiers, as well as supplementary geographic characteristics, are available for the person's residence and workplace, thereby allowing for a detailed spatial analyses. Austria has exceptionally comprehensive data on commuting that has been instrumental in various studies, including for research on tax evasion (Frimmel et al., 2019). The Employment Statistics encompass detailed information about the commuting patterns of Austrian workers, capturing both the distance travelled, measured in kilometres, and the travel time, recorded in minutes, between their place of residence and their workplace. Such granularity in the data enables an

[6] The current data catalogue can be found at https://www.statistik.at/amdc-data/#/product

[7] See https://www.statistik.at/en/about-us/surveys/surveys-a-z for a list of all surveys.

in-depth analysis of commuting behaviours and patterns across different regions of Austria.

The **Education Register** (*Bildungsstandregister*) is integrated into the Employment Statistics and tracks the educational attainment of all Austrian citizens from primary school to the highest type of education. Like the Employment Statistics, it is available since 2009 and is always measured annually at the end of October. The register reports both current enrolment in education and completed levels of education. The field of study is grouped following the International Standard Classification of Education (ISCED) in order to facilitate cross-country comparisons. The dataset also provides spatial information on the respective school location, the distance from home and the corresponding travel time. The data's longitudinal nature allows researchers to track educational progress over time by providing insights into the effectiveness of educational policies and programs. Moreover, the integration of geographic information allows for the study of regional disparities in educational attainment. Policymakers and researchers can better understand and address the educational needs and challenges within the Austrian context by leveraging the comprehensive data available in the AMDC.

As a complement of the annual labour force assessment of the Employment Statistics, the **Register-Based Employment Spells** (Registerbasierte Erwerbsverläufe) provides spell-level data about the work history of all workers since 2010. The data encompass both employment and unemployment spells, given that it is compiled from administrative records of the Social Security Administration and the Public Employment Service (AMS), and this allows for more-fine grained labour market analyses as the employment status of workers can be observed at any given point in time.

The combination of several datasets in the AMDC allows for family linkages between spouses and between parents and children to be uncovered. Marital and birth records are available for the whole population for recent family links. The **Marital and Birth Register** (*Statistik der Standefälle*) has recorded the dates of all marriages, births and deaths since 2015. Divorce records are available in a separate database (*Statistik der Ehescheidungen*) starting in 2018. The person identifiers in these registers allow for the matching of married spouses as well as parents with their children. Earlier years of data are not available (yet), unfortunately, because the corresponding person identifiers have not been assigned for these records. The AMDC also reconstructs parent-child links for previous periods, dating back to cohorts of children born in the 1980s, by using a combination of coinsurance records and population registers. These linkages enable researchers to exploit AMDC data for the analysis of intergenerational mobility.

The AMDC also includes data from various tax registers for every fiscal year since 2000. The **Income Tax Statistics** and **Wage Tax Statistics** provides records for all Austrian citizens who are subject to taxation. In addition to data on taxable income, tax allowances and taxes paid, the registers also include variables on benefit payments such as childcare allowances. Moreover, the AMDC contains similarly structured datasets on corporate taxes and sales taxes, thereby substantially extending the scope for firm-level analyses.

3 Austrian Labor Market Database (AMDB)

While the AMDC was just recently established, matched employer-employee records have long been used for policy and academic research in Austria. Zweimüller et al. (2009) provide a comprehensive description of the data and its various components. Most of the data are accessible to researchers via the Austrian Labor Market Database (AMDB, Arbeitsmarktdatenbank), which is jointly offered by the Federal Ministry of Labour and Economy and the Public Employment Service AMS. The AMDB can be downloaded (for a yearly subscription fee) provided that subscribers confirm that the data can be hosted within a secure research infrastructure. Similarly to the AMDC, access is not restricted to Austria, but researchers have to sign a confidentiality agreement and submit a project proposal with their data application.[8] While data updates in the AMDC often come with year-long delays, most of the AMDB datasets are updated on a monthly basis. This enables researchers to closely track recent changes in the labour market; this can help policy makers to receive real-time policy recommendations shortly after the implementation of new reforms. However, external data cannot be merged to microdata from the labour market database.

3.1 Overview of Datasets

The AMDB contains various datasets that, when used in conjunction, allow for comprehensive coverage of the Austrian labour market.[9] Some of these data are now also accessible in the AMDC, but the AMDB provides more detailed information on the worker's employment history for a much longer time span. At the core of the database are **spell-level data from the Social Security Administration** with precise start and end dates, which are available from 1997 onwards. In addition to the employment spells of private sector workers in Austria, the dataset also includes many other spell types often overseen by the social security administration, such as spells related to (long-term) sick leave, pension, disability or parental leave.[10] Data on parental leave episodes, for instance, have been used to study the effects of changes in benefit schemes on leave-taking choices (Ziegler & Bamieh, 2023b); child penalties and parental leave policies (Kleven et al., 2024) and the relative importance of job protection and cash benefits in parental leave systems and for mothers' careers (Lalive et al., 2014). The social-security spells are supplemented with **spell-level data from the Public Employment Service AMS** that cover all spells of registered jobseekers since 1985, including benefit receipt and participation in AMS training schemes. Importantly, both workers and firms are assigned anonymous and unique identifiers

[8] Details about the application procedure and the application form can be found here: https://arbeitsmarktdatenbank.at/

[9] A detailed description of datasets is available online (https://arbeitsmarktdatenbank.at/docs/Beschreibung_AM-Datenbank-Tabellen%20Uni_Status.pdf)

[10] Public sector workers are only observed in more recent years.

in order to track them over different spells. Supplementary datasets contain the individual characteristics of all workers (gender, age and nationality) and firms (sector and region). The combination of work and unemployment spells allow researchers to study various types of labour market dynamics such as job search behaviour (Card et al., 2007).

Another database reports every worker-firm-link's monthly labour earnings up to the maximum contribution threshold that applies to social security payments (in 2024 this was €6060 per month). Other income components such as unemployment benefits or parental leave allowances can be inferred from these records. The earnings data have, for instance, been used in the literature to study the effect of unemployment insurance on reemployment wages (Nekoei & Weber, 2017).

Finally, the AMDB also contains coinsurance spells and birth records. The spells enable linkages of parents to children and of parents to each other because children are covered by health insurance through their parents. These data can, for instance, be used to study family dynamics and differences between spouses (Ziegler & Bamieh, 2023a). The data on newborns allow for fertility patterns to be studied; Lalive and Zweimüller (2009), for example, examine how changes in the duration of paid, job-protected parental leave affect mothers' fertility.

4 Conclusion

While access to register data had been limited in the past, the Austrian Microdata Center (AMDC) and the Austrian Labor Market Database (AMDB) have significantly expanded opportunities for empirical researchers. A key novelty of these initiatives is that linkages across registers from different sources are now feasible without the necessity to seek approval from multiple data providers, thereby opening up new avenues for interdisciplinary research.

Future efforts ought to be directed towards expanding the stock of databases that can be both readily linked and analysed within the Austrian Microdata Center. External data providers, such as ministries or government agencies, should facilitate the merging and use of additional registers within this infrastructure. Given the AMDC's high data protection standards and output control, researchers can now analyse these data without compromising the confidentiality of individual-level records. While certain domains are already broadly represented in the AMDC, such as in the labour market and education, the stock of databases in other areas, such as health, could still be expanded.

Another limitation is that many registers do not extend far into the past, which can be explained by the lack of unified identifiers in previous years. This constrains the analysis of long run trends, intergenerational connections or the evaluation of policies in the more distant past. The respective data providers should also aim to create identifiers for earlier years in order to further extend coverage of the registers into the past. Ideally, there will be one unified platform in the future, thereby

offering researchers both affordable and comprehensive access to the universe of administrative data in Austria.

List of Useful Links

- Statistics Austria: https://www.statistik.at/en/
  - List of surveys: https://www.statistik.at/en/about-us/surveys/surveys-a-z
- Austrian Micro Data Center (AMDC): https://www.statistik.at/en/services/tools/services/center-for-science/austrian-micro-data-center-amdc
  - Data catalogue: https://www.statistik.at/amdc-data/#/product (in German).
  - Cost catalogue: https://www.statistik.at/fileadmin/pages/1805/Katalog_der_Serviceleistungen.pdf (in German).
- Austrian Labor Market Database (AMDB): https://arbeitsmarktdatenbank.at (in German).
  - Description of datasets: https://arbeitsmarktdatenbank.at/docs/Beschreibung_AM-Datenbank-Tabellen%20Uni_Status.pdf (in German).

References

Abowd, J. M., Kramarz, F., & Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica, 67*(2), 251–333.

Card, D., Chetty, R., & Weber, A. (2007). Cash-on-hand and competing models of intertemporal behavior: New evidence from the labor market. *Quarterly Journal of Economics, 122*(4), 1511–1560.

Frimmel, W., Halla, M., & Paetzold, J. (2019). The intergenerational causal effect of tax evasion: Evidence from the commuter tax allowance in Austria. *Journal of the European Economic Association, 17*(6), 1843–1880.

Kleven, H., Landais, C., Posch, J., Steinhauer, A., & Zweimüller, J. (2024). Do family policies reduce gender inequality? Evidence from 60 years of policy experimentation. *American Economic Journal: Economic Policy, 16*(2), 110–149.

Lalive, R., Schlosser, A., Steinhauer, A., & Zweimüller, J. (2014). Parental leave and mothers' careers: The relative importance of job protection and cash benefits. *Review of Economic Studies, 81*(1), 219–265.

Lalive, R., & Zweimüller, J. (2009). How does parental leave affect fertility and return to work? Evidence from two natural experiments. *Quarterly Journal of Economics, 124*(3), 1363–1402.

Nekoei, A., & Weber, A. (2017). Does extending unemployment benefits improve job quality? *American Economic Review, 107*(2), 527–561.

Ziegler, Lennart, & Omar Bamieh. (2023a). *What drives paternity leave: Financial incentives or flexibility?* IZA Discussion Papers, No. 15890.

Ziegler, Lennart, & Omar Bamieh. (2023b). *Does a flexible parental leave system stimulate maternal employment?* IZA Discussion Papers, No. 16172.

Zweimüller, J., Winter-Ebmer, R., Lalive, R., Kuhn, A., Wuellrich, J. P., Ruf, O., & Büchi, S. (2009). *Austrian social security database.* Available at SSRN 1399350.

Omar Bamieh is an Associate Professor at the Institute for Economic Policies, Universidad Andrés Bello (Santiago, Chile) and a research fellow at IZA (Institute of Labor Economics). Prior to this appointment, he was an assistant professor of economics at the University of Vienna. He studied at Bocconi University (Italy) and the European University Institute (Italy) where he completed a PhD in economics. His research interests include Labor Economics, Occupational Licensing, Gender Pay Gap, Gender Discrimination, Parental Leave, Social Norms, Wage Transparency, and Work from Home (WFH). In recent years, he has worked with both the Austrian Micro Data Center and the Austrian Labor Market Database.

Lennart Ziegler is an assistant professor in economics at the University of Vienna and a research fellow at IZA (Institute of Labor Economics). Prior to his current position, he studied at the University of Mannheim (Germany) and the Tinbergen Institute (Netherlands) before completing a joint doctorate at the University of Amsterdam and VU Amsterdam. His research agenda covers various topics such as unemployment, intergenerational mobility and gender differences. Lennart has gained extensive experience with both the Austrian Micro Data Center and the Austrian Labor Market Database. He teaches courses in Labor Economics and Applied Microeconometrics to Bachelor and Master students at the University of Vienna.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 3
The Use of Administrative Data for Evidence-Informed Policy Development in Belgium



Koen Vleminckx

Abstract This chapter highlights Belgium's long-standing commitment to leveraging administrative data for evidence-informed policymaking. It discusses relevant particularities, regarding the implementation of the GDPR in Belgium, and examines the roles played by both the Data Protection Authority (DPA) and the Information Security Committee (ISC) in ensuring data protection and compliance. It also details how data are made available to government experts and to the scientific community via Belgium's statistical office, Statbel, and the Datawarehouse Labour Market and Social Protection (DWH LM&SP) of the Crossroads Bank of Social Security (CBSS). This chapter then presents some interesting initiatives specifically aimed at unlocking administrative data for statistical monitoring, policy evaluation, evidence-informed policy making and scientific research. The document underscores the importance of collaboration between regulatory authorities, researchers and policymakers in order to balance innovation and accountability in data use.

1 Introduction

The need for data for evidence-informed policy making and policy evaluation is not new. In Belgium, this need was expressed as early as the 1980s by policy makers that wanted to be able to use robust and timely evidence to guide the country, and the Belgian welfare state in particular, through years of economic crisis. This acute need has lead to successful initiatives, such as the creation of the Crossroads Bank for Social Security (CBSS), and regulatory measures to ensure privacy and data security, while allowing for the use of data for both policy development and scientific research.

Belgium's federal and regional administrations make aggregated data available at a fairly detailed level within the context of 'open data' and they each have an open data portal. The federal (Statbel) and regional statistical offices also have their own open data portal and Statbel makes aggregated data available at a detailed level.

K. Vleminckx (✉)
FPS Social Security, Brussels, Belgium
e-mail: Koen.Vleminckx@minsoc.fed.be

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_3

However, we will focus on initiatives that make administrative data accessible for evidence-informed policy making, policy evaluation and scientific research in this chapter. As the author works in the domain of social protection, there is a delimitation, or bias if you like, concerning initiatives in this policy domain.

2 Regulatory Framework on the Use of Administrative Data for Scientific Research and Evidence-Informed Policy Making

In Belgium, the General Data Protection Regulation (GDPR) of 27 April 2016 is implemented through a set of federal and regional regulations.[1] In doing so, several legislative initiatives have extended the protection sought at the European level. The most important of these is the federal Act of 30 July 2018 concerning the protection of natural persons with regard to the processing of personal data. This Act has, depending on the purpose of data use, built in additional safeguards and/or procedures.

Belgian regulations include provisions on the rights of individuals whose data are processed and also specify the circumstances under which those rights can be limited. The Belgian Act of 30 July 2018 details in which cases and to which processors/sources the exception applies in the event that the GDPR has left room for Member States to determine exceptions to parts of the protection. The Belgian Act does not grant discretionary power to the executive (e.g., to determine modalities for processing data for entities in the public administration).

Belgium is one of the few EU Member States to have a GDPR research regulation, namely the Title IV of the Act, which concerns processing for the purposes of archiving in the public interest, scientific or historical research or statistical purposes referred to in Article 89 §§ 2 and 3 of the GDPR. Some Belgian stakeholders sought a framework that would provide them with more legal certainty. More specifically, this

[1] At the federal level:

- Act of 30 July 2018 on the protection of natural persons with regard to the processing of personal data;
- Act of 3 December 2017 establishing the Data Protection Authority (DPA);
- Act of 5 September 2018 establishing the Information Security Committee (ISC).

At the Flemish level:

- Decree on electronic administrative data traffic (E-Gov Decree);
- GDPR decree of 8 June 2018 reflects, among other things, the establishment of the Flemish Supervisory Commission (VTC);
- A BVR dated 19 July 2019 will show for which domains where personal data is used one can file an appeal with the VTC.

Title IV defines the conditions for derogations regarding the rights of data subjects referred to in Article 89 §§ 2 and 3 of the GDPR.

Furthermore, various bodies were established by Acts to implement the regulation, namely the Data Protection Authority[2] and the Information Security Committee.[3] These Acts both play a role in data protection and compliance with privacy and information security legislation.

The Data Protection Authority (DPA) is the Belgian supervisory authority for the protection of personal data. It monitors compliance with the GDPR and Belgian privacy legislation. The DPA is the successor to the Commission for the Protection of Privacy, so-called 'Privacy Commission,' established under the 1992 Privacy Act and updated alongside the GDPR. Its functions are summarised as follows:

- Monitor and enforce compliance with the GDPR;
- Handle complaints from citizens about possible violations of their privacy rights;
- Advising organisations and governments on privacy issues;
- Impose fines or penalties in case of privacy violations.

The Information Security Committee (ISC) is an inter-federal committee that focuses on the security and use of sensitive data, mainly within the public sector. The ISC has specific tasks in the field of information security–including the granting of deliberations for certain types of personal data communications–but it is not a supervisory authority within the meaning of the GDPR. Consequently, the ISC is not competent for monitoring compliance, resolving problems and disputes or for handling complaints. Indeed, the DPA is competent in such instances.

The DPA has a broader role to play, overseeing all data processors in Belgium (both public and private), while the ISC focuses more specifically on information security and data exchange within the public sector. There is some overlap between the activities of both bodies, especially when it comes to GDPR compliance and personal data security. It is possible for the two bodies to cooperate in cases in which information security and privacy protection intersect. The DPA may consider advice from the ISC when assessing complex issues in the public sector.

The DPA ensures that researchers and bodies wishing to use administrative data comply with the rules of the GDPR regarding the use of administrative data for scientific research. This means, among other things, that data processing must be lawful (a legal obligation, public interest, or some such) and the use is done in accordance with the principles of purpose limitation, data minimisation and security. The DPA can also advise investigation teams on how to process data in a compliant manner.

Thus, the DPA acts as a regulator of the overall process and intervenes when there may be violations of privacy laws. It monitors compliance with GDPR rules at a higher legal level and ensures that processing is carried out in accordance with privacy legislation.

---

[2] Act of 3 December 2017 establishing the Data Protection Authority (DPA).

[3] Act of 5 September 2018 establishing the Information Security Committee (ISC).

The ISC plays a key role in approving requests made for access to administrative data, particularly regarding data originating from public health and social security databases (such as the CBSS, eHealth platforms, etc.). The ISC assesses whether the data processing is justified and proportionate, and whether information security requirements have been met. The ISC may impose conditions on the use of data, such as pseudonymising or anonymising data, or restricting access to a secure environment. The ISC then approves data sharing, provided that all legal and security requirements are met.

The ISC has two chambers: the Federal Government Chamber and the Social Security and Health Chamber. The ISC's Federal Government Chamber deliberates on the communications of personal data by government departments and public institutions of the federal government to third parties that are not social security institutions, to the extent that the controllers concerned do not reach an agreement or at least one controller requests a deliberation. In contrast to the ISC's Social Security and Health Chamber, the Federal Government Chamber rarely intervenes in the case of data requests for research purposes, because a contract is usually concluded between the researcher and a public administration in the case of non-sensitive data.

The Social Security and Health Chamber is the successor to the Social Security and Health Sectoral Committee, from which it has taken over some of its competences, such as formulating good practices, granting consultations and supporting data protection officers. The ISC's Social Security and Health Chamber is responsible for assessing and approving applications relating to the exchange and processing of administrative data in the fields of social security and health. This chamber, thus, handles data exchanges that often contain sensitive information, such as medical records, health data, social security information (e.g., unemployment, pension) and other personal data. The regulatory framework within which this chamber operates includes, in addition to GDPR and the Act of 30 July 2018, the Act of 5 September 2018 on its establishment, the Act of 21 August 2008 establishing and organising the eHealth Platform and the Act of 15 January 1990 establishing and organising a Crossroads Bank of Social Security (CBSS).

3 Using Administrative Data for Scientific Research and Evidence-Informed Policy Making

Between 1998 and 2011, the AGORA programme funded projects relating to the development, exploitation and creation of either administrative or non-administrative data.[4] AGORA has played a crucial role in making Belgian federal government administrative data accessible for scientific research and policy development, monitoring and evaluation. AGORA also facilitated the development and use of existing

[4] The Agora programme was overseen by an Interdepartmental Coordination Committee (ICC) comprised of representatives mandated by federal institutions and chaired by the Office for Scientific, Technical and Cultural Affairs (DWTC), now called the Federal Science Policy Office (BELSPO).

administrative databases within federal government departments. This enabled researchers to access valuable data for in-depth scientific research and supported policy makers in formulating and evaluating policies based on empirical data.[5]

Through these initiatives, AGORA has contributed to improved access to and use of administrative data from many policy domains for research purposes, including scientific support for policy, which in turn has led to more well-informed policy decisions and more effective monitoring and evaluation of policy measures in Belgium.

Numerous Belgian federal and regional administrations and institutions have, of course, also made their own direct investments in recent decades in opening up the administrative data that they manage for statistical monitoring, policy evaluation and preparation and for scientific research.

The sections below describe some notable initiatives, beginning with Belgium's Central Statistical Office, Statbel and followed by the Datawarehouse Labour Market and Social Protection (DWH LM&SP) of the Crossroad Bank of Social Security (CBSS), which is an essential source of administrative data for both government experts and scientists engaged in evidence-informed policy making, as well as for the broader scientific community. Furthermore, we present interesting initiatives that are specifically aimed at unlocking administrative data for statistical monitoring, policy evaluation, evidence-informed policy making and scientific research:

- Dynam: Labour market dynamics;
- Permanent Health Care Sample;
- Health Data Authority;
- PensionStat.be.

3.1 Statbel

Belgium's statistical office, Statbel, manages a wide range of microdata derived from administrative sources, such as data on demography, labour, housing, income and social security. In order to compile its statistics, Statbel uses a legal mandate to receive administrative databases and in practice uses existing administrative databases as much as possible.[6] Administrative data collected by Statbel are immediately pseudonymised. Statbel itself stores the key for pseudonymisation to be able to link data thereafter.

Statbel's processing of data and the sharing of this data with third parties is regulated by law through the Law on Public Statistics of 4 July 1962 (hereinafter "Statistics Law"). The basic principle of the Statistics Act is that all data obtained

[5] The second component was the creation of new data sources; in addition to exploiting existing data, AGORA also supported the collection of new data by researchers themselves. This included, for example, conducting ad hoc surveys to fill in gaps in existing data, so as to get a more complete picture of social issues.

[6] On its website, Statbel publishes a list of the administrative databases that it uses:https://statbel.fgov.be/en/about-statbel/privacy/privacy-gdpr.

and processed by Statbel for statistical purposes are protected by statistical secrecy (both data about natural persons and legal entities). Furthermore, as regards personal data, the GDPR and the Act of 30 July 2018 on the protection of natural persons regarding the processing of personal data also apply.

Statbel makes pseudonymised data available for research and evidence-informed policy making, including consolidated data from administrative sources used by Statbel. A communication of pseudonymised data with a limited risk of indirect identification is possible (through a confidentiality agreement), provided that the researchers themselves do not have access to the additional data required for indirect identification.

Researchers can request microdata if the data is necessary for their statistical or scientific research. "Necessary" means that the research is not possible on the basis of global and anonymous data that Statbel disseminates without restrictions via the website https://statbel.fgov.be/en.[7] Statbel makes aggregated data available at a detailed level through Open data; the purpose for which this data is requested is limited to statistical and scientific research. Policy preparatory research or research by private institutions are also eligible, provided that the processing happens independently, transparently and by making use of scientific methods. The results of the research should also be made public.

The result of the processing must consist of global and anonymous statistics and research reports that may not have any individual impact on either the citizens or enterprises concerned. Of course, these published global and anonymous results can be freely used afterwards for non-statistical or non-scientific purposes.

Given the nature of the data, the procedure is limited to a recommendation from Statbel's Data Protection Officer (DPO) and the approbation of the Director-General of Statbel in most cases.

European[8] research centres from both universities and multilateral organisations also have access to pseudonymised data. In order to gain access, they must first apply to Statbel using a standardised application form. Based on the form, a contract is drawn up and is presented to Statbel's DPO for advice. Statbel's Director-General must then grant authorisation. After the contract has been signed, the pseudonymised data will be delivered encrypted via a secure connection.

Statbel also makes aggregated, anonymous data available as 'open data' via the Statbel website. Aggregated data with a limited risk of indirect identification are available via a confidentiality contract.

Statbel provides documentation and statistical information in English. The English version of their website is accessible via statbel.fgov.be/en, where you can access various topics such as population, economy, labour market, agriculture, energy and more.

Open data provided via the Statbel Open Data portal is freely available to anyone, including third parties in other EU member states, without charge. Statbel may

[7] Be. STAT, Statbel's electronic database, also allows researchers to create customised tables.

[8] For researchers based at universities outside the EU, it is verified whether or not they have an agreement with the EU regarding compliance with privacy.

charge a fee for customised or specialized data requests (tailored datasets, analyses or services not available directly through open platforms). Such charges typically reflect the costs associated with preparation, processing and delivery of data.

3.2 The Crossroads Bank for Social Security's Datawarehouse Labour Market and Social Protection (DWH LM&SP)

The first project funded under the AGORA programme (1999, AG/DD/001), provided for the creation of the Datawarehouse Labour Market, now known as the Datawarehouse Labour Market and Social Protection (DWH LM&SP), of the Crossroads Bank for Social Security (CBSS).[9] This Datawarehouse has become an essential source of administrative data for both government experts and scientists engaged in evidence-informed policy making, as well as for the broader scientific community, over the past quarter of a century.

This project's aim to set up a system within the CBSS to store anonymised, processed and harmonised data from partner institutions. The aim was to use this DWH LM&SP to provide a more flexible, rapid, and cost-effective response to users' requests for statistical data that used to be addressed to the CBSS's partner institutions.

The DWH LM&SP aims to link socio-economic data that initially came mainly from the operational databases of public social security institutions, but has been expanded to include data from many other federal and regional public institutions that have further enhanced the relevance of the DWH LM&SP for scientific research, policy preparation, monitoring and evaluation over the past decades.

The DWH LM&SP contains a wide range of data, including:

- Personal characteristics: Information such as age, gender and place of residence;
- Education and qualifications: Data on level of education and qualifications obtained;
- Labour market situation: Details on employment, unemployment, and occupational status;
- Income data: Information on diverse types of income, such as wages and benefits;

[9] This came at the request of the CBSS, in implementation of Article 5 of the Act of 15 January 1990 establishing and organising a CBSS, and from a number of other public social security institutions. A major motivation for the creation and organisation of the CBSS Act of 15 January 1990 was to improve administrative efficiency, both for public administrations and Belgian citizens, by creating a repository of basic data obtained from the socially insured persons, so that the citizens and the bodies concerned were relieved of superfluous administrative formalities. However, the CBSS was also conceived of as a policy support instrument. Indeed, through the aforementioned network, "a quasi-inexhaustible arsenal of data" is made available "which can be processed into important statistical, sociological, economic and financial information for the benefit of the various management and decision-making bodies" (Robben, Viaene, 1989: 161).

- Social benefits: Data on child benefits, pensions and other social allowances received.
- Health-related data: Disability and disability status information.

The DWH LM&SP has become an important data source for government experts, statisticians, and scientific researchers alike over the past decades, not only in the field of labour market and social protection policy, but also in many other disciplines and policy domains. This is not least because, subject to the necessary permissions, pseudonymised data from the DWH LM&SP can also be linked to other data from other data sources that are not included in the DWH LM&SP.[10] This provides a way to acquire pseudonymised administrative data on the population's social and economic characteristics, even for small target groups, that can be coupled with other datasets (even survey data) using a unique identification number.

As a result, it has become the main gateway to obtain such administrative data and the DWH LM&SP is used for various purposes; interesting examples in the domain of policy monitoring and evidence-informed policy making include:

- Labour market report from the 'Steunpunt Werk':

  The 'Steunpunt Werk' (which can be translated as 'Support Point Work') uses the DWH LM&SP to analyse labour market trends, on behalf of the Flemish Government. These data support regular monitoring reports and specific studies. A new dashboard on labour market slack will also be launched In 2025, as well as a new Flemish Labour Account with local labour market statistics down to the level of individual municipalities.

  More information: Steunpunt Werk, https://www.steunpuntwerk.be/
- Belgian Microsimulation Model on Social Security and Taxes (BELMOD).

  BELMOD is the FPS Social Security's tax-benefit microsimulation model that allows for policy changes in social benefits and taxes at both the household and individual levels to be simulated. BELMOD demonstrates the added value of using administrative data for ex ante distributional impact analysis. The BELMOD model allows modelling not only broad tax and benefit reforms, but also specific measures or measures that are targeted to extremely specific groups by making use of the detail that is available in administrative data for a 1.4 million sample of the population. BELMOD estimates the budgetary implications and the overall effect on poverty and income inequality, as well as the effect on specific groups that are usually difficult to identify in survey research. For example, BELMOD has been used to show that small increases in categorical income support for people with disabilities have only a limited impact on the overall poverty rate, but can significantly reduce the poverty among the target population.

  BELMOD is also a valuable tool in the cooperation between national government administrations, academics and international research institutions. This is facilitated by the fact that BELMOD is embedded in the EUROMOD infrastructure, the European Commission's tax-benefit microsimulation model that is

[10] However, this requires the necessary methodological expertise from the scientific user, as well as the necessary time to go through and to fulfil the procedural requirements.

developed and maintained at the JRC. This enables cooperation with the JRC (e.g., for the purposes of linking BELMOD with EUROLAB). EUROLAB is a behavioural microsimulation model that is used to assess the impact of fiscal or economic policies on labour supply and the overall employment in EU countries. Moreover, BELMOD is accessible to external researchers via a remote access tool. This also enables collaboration with academic teams. BELMOD is used in this way by the University of Antwerp to refine the measurement of inactivity traps in Belgium and by the KU Leuven to map the distributional effects of carbon taxes.

More information about BELMOD in English, including technical documentation, is available on the dedicated web pages of the FPS Social Security: https://socialsecurity.belgium.be/en/shaping-social-policy/belmod

- Socio-economic monitoring report 'Labour Market and Origin' (UNIA in collaboration with FPS Employment).

  The objective of this monitoring report is to provide insight into the labour market position of persons with a migration background, including indicators such as employment, unemployment and wage equality. Using the data warehouse data can reveal differences between groups and can contribute to more inclusive policies.

  More information: UNIA, https://employment.belgium.be/en/publications/5th-socio-economic-monitoring-labour-market-and-origin-2022

The CBSS distinguishes between two types of applications regarding the accessibility of DWH LM&SP data:

- Web applications: these applications allow users to consult statistics about the population in Belgium and persons abroad with a link to Belgian social security and provide insight into variables, such as socio-economic position, place of residence and employment;
- Customised data applications: these are specific applications requested by research institutes and government departments.

Regarding customised data requests: the data provided by the CBSS may only be used for research purposes and then only for the probable duration of the research, which must also be communicated in advance. After the probable duration of the research or once the research purposes have been completed, the organisation must destroy the data communicated. It may, however, request that the CBSS continue to hold this data for a certain period, for justification of the research results or for follow-up research. Under no circumstances may data communicated for research purposes be passed on to third parties. This principle can only be deviated from if the requesting organisation both demonstrates and justifies the necessity of passing on the data to third parties in function of the specified research in its application.

A distinction should further be made between anonymous data, pseudonymised personal data and non-pseudonymised personal data. The CBSS operates a "tiered" system, so as to respect the principle of minimum data processing. Research should, in principle, be done with anonymous data. Pseudonymised personal data may only be used provided that a processing of anonymous data does not make it possible to

achieve the research purposes. Non-pseudonymised personal data can only be used provided that the processing of pseudonymised personal data also does not make it possible to achieve the purposes of the investigation.

3.2.1 Application to Obtain Anonymous Data

If an organisation wishes to use anonymous data from a single body for research purposes, then it must make a request to that body itself, which will decide whether or not to grant the request; it is then also responsible for anonymising and communicating the personal data itself. The CBSS does not have to intervene in such instances. If an organisation wishes to use anonymous data from various sources for research purposes, then it must submit a request to that effect to the CBSS, which will decide whether or not to grant the request.

3.2.2 Request to Obtain Pseudonymised Personal Data

Pseudonymised personal data are data that can only be linked to identified or identifiable natural persons by means of a code; the organisation processing them for research purposes cannot itself trace them back to the natural persons concerned. The link with the data subjects is kept outside the organisation (e.g., the CBSS itself), such as for follow-up research or for a justification of the research. If an organisation wishes to use pseudonymised personal data from various sources for research purposes, then it must submit an application to the CBSS, which will decide whether to grant the request. In that application, the organisation must thoroughly demonstrate that the personal data is indeed pseudonymised and that the persons concerned cannot reasonably be re-identified based on the information requested. However, it is not enough to simply remove the identification number, surname, first name and address of the person concerned to be able to speak to the pseudonymised personal data: their characteristics being too specific must be avoided. Thus, the data provided must be sufficiently generalised, by using sufficiently broad value classes for example, and the number of selection criteria to be combined must not be so large that the number of persons meeting them is exceedingly small and those persons, therefore, become re-identifiable.

If an organisation wishes to use pseudonymised personal data from a single social security institution for research purposes, then it must submit an application to that social security institution itself, which will decide whether to grant the request. However, the social security and health chamber of the information security committee must also grant an approbation. If the information security committee grants an approbation, then the social security institution processes its personal data

and communicates them in a pseudonymised fashion. In principle, the CBSS does not then have to intervene in such cases.[11]

3.2.3 Informing the CBSS

The CBSS itself does not have to intervene when one of the sources communicates data to an organisation for research purposes, then it must always inform the CBSS in one way or another, even if for applications relating to the aforementioned principles. The CBSS, thus, retains an overview of the various communications and is able to prevent data links from being carried out, contrary to what actions that precede by bodies other than itself.

3.2.4 Accessing the Data

The disclosure of pseudonymised personal data by the CBSS may be done in two phases, in order to ensure better protection of the privacy of the socially insured persons involved. In the first phase, the CBSS shares pseudonymised personal data from a sample of the target population with the researchers for the development of algorithms. In the second phase, the researchers have access to the same pseudonymised personal data from the entire target population–in the CBSS building, on a secure computer and under permanent supervision–to apply their previously created algorithms and they can only bring the results of their actions in the form of purely anonymous information outside the CBSS building (the CBSS carries out a small cell risk analysis in advance for this purpose).

The CBSS is currently exploring whether innovative technologies can be used to access data more efficiently, such as via a remote access system. In this context, it is worth mentioning that the FPS Social Security, in cooperation with Sciensano, has developed a Remote Access platform that allows partners to consult and use the BELMOD model and even underlying data via a secure connection. BELMOD, thus, becomes accessible to the wider research community. The use of Remote Access involves some conditions in order meet security requirements.

[11] However, the social security institution concerned may also decide to entrust the processing of its personal data to the Crossroads Bank, given the latter's experience and expertise as an intermediary organisation. In that case, it forwards the application to the Crossroads Bank, which processes, pseudonymises and communicates the personal data after deliberation by the Social Security and Health Chamber of the Information Security Committee.

3.2.5 Circle of Trust[12]

In order to be considered a Circle of Trust,[13] a research organisation must meet the eleven criteria listed in the regulations establishing the criteria for the application of a circle of trust in the processing of personal data in the context of investigations useful for the knowledge, conception and management of social protection, approved by the information security committee. The research organisation must take all necessary measures to ensure compliance with these criteria. The establishment of the Circle of Trust is accompanied by a declaration of honour from the research organisation, in which it explicitly declares that it complies with the eleven criteria that comprise the standard framework. By establishing a Circle of Trust, a data application from a research organisation can be handled more efficiently by the information security committee. When a Circle of Trust has been established, the Information Security Committee no longer needs to examine the substance of the security measures taken because it can be assumed that the eleven criteria have been met.

The CBSS provides detailed information about the Datawarehouse Labour Market and Social Protection Data on its website, but only in French and Dutch: https://dwh.ksz-bcss.fgov.be/fr/homepage/index.html

The BCSS may charge a fee for customised or specialised data requests (tailored datasets not available directly through open platforms). Such charges typically reflect the costs associated with preparation, processing and delivery of data.

3.3 Dynam: Labour Market Dynamics

Dynam was created in 2010 with the purpose of mapping the dynamic dimension of the labour market in Belgium. It was initially a collaboration between Higher Institute of Labour (HIVA) of the KU Leuven and the National Social Security Office (NSSO), and with the support of the Federgon Chair 'Dynamics of the Labour Market' (KU Leuven) between 2001 and 2014. In 2014, this cooperation was extended to the regional entities (Brussels: BISA, Flanders: Department WSE, and Wallonia: IWEPS).

Over a period of almost 15 years, the Dynam project grew into a unique source for figures about the dynamics of and in the Belgian labour market, both at the federal and regional levels and at the subregional level of provinces and even in smaller reference regions, based on administrative data. The focus was initially on visualising gross job

[12] Information about the Circles-of-Trust are available in Dutch and French on the dedicated web pages of the CBSS: https://dwh.ksz-bcss.fgov.be/fr/dwh_page/content/websites/datawarehouse/menu/cercle-de-confiance.html

[13] More information: https://dwh.ksz-bcss.fgov.be/fr/dwh_page/content/websites/datawarehouse/menu/cercle-de-confiance.html

creation and the destruction in the labour market using, corrected, employee flows captured by NSSO administrative data.[14] The annual series contains the evolutions as of 30 June for each year, and this covers over 99% of wage-earning employees in Belgium. A further deepening followed in 2013 with the addition of worker dynamics (entry and exit movements of workers). Only two years later, the cooperation with and input of the regional entities made it possible to map regional and subregional.

A recent development concerning the link between Dynam and the aforementioned DWH AM&SB is that other profile characteristics from other administrative databases can also be linked to worker dynamics figures: nationality, education level, socio-economic position and others.

Dynam is an inter-federal platform and acts as a hub for developing, producing and publishing statistics on labour market dynamics in a way that is scientifically supported by the research team at HIVA-KU Leuven. In 2018, it invested in a new website (www.dynamstat.be), which is fully accessible in two languages. Dynam's website provides an interactive and dynamic environment to retrieve the most important sets of figures. With this opening-up of the data, Dynam meets the increasing need of users to consult the data themselves and to customise the data request (subregional level, comparison over time, detailed subsectors, etc.).

More and more users (government bodies, social partners, employers' federations, etc.) are finding their way to the Dynam-Reg figures, often with a second movement of demand for deeper customised analyses (evolution of inflow and outflow dynamics, employee profile, origin of new employees, length of stay in new job, etc.). Moreover, Dynam is frequently used by policy bodies, for monitoring and for research purposes. In scientific circles, the importance of documented and corrected administrative data with evolution over several measuring points is also growing.

The Dynam consortium provides detailed information (methodology, statistics, publications) about the Dynam project on its website, but only in French and Dutch: https://www.dynamstat.be/

3.4 Permanent Health Care Sample

The 'Permanent Sample' is a dataset on healthcare utilisation in Belgium. The Permanent Sample (EPS) is a longitudinal dataset containing administrative data from the Population, Health Care and Farmanet databases, as well as data on hospitalisations as in the Hospitalisations database. The Intermutual Agency (IMA)[15] creates the sample database.

[14] This includes corrections for apparent changes in employment due to demergers, acquisitions, mergers or administrative changes.

[15] More information on IMA-AIM: https://www.ima-aim.be/?lang=nl

These data do not cover the total population, but only an anonymous, representative selection of 300,000 people from those affiliated to a Belgian health insurance fund for their compulsory health care insurance. More people aged sixty-five and over were included in order to better study the effects of population ageing In this representative panel sample by age and sex.[16] For these (anonymous) individuals, the following data relevant to the health care provided to them are then included:

- consultations and visits to doctors;
- prescribed drugs;
- hospitalisations;
- nursing performance;
- etc.

Derived variables are also created and made available in the database.

The anonymised data are available in separate datasets by calendar year.

Documentation for the Permanent Health-Care Sample is available in both Dutch and French. The IMA provides detailed metadata and descriptions of the dataset on their website in both languages: https://metadata.ima-aim.be/fr/app/bdds/Ps

Access to the Permanent Health-care Sample is restricted, given the sensitivity of the data, even though it is anonymous,. Currently, the Permanent Sample is used by experts from the NIHDI, the FPS Public Health, Food Chain Safety and Environment, the FPS Social Security, Sciensano, the Federal Knowledge Centre for Health Care (KCE), the Intermutual Agency (IMA), the Federal Planning Bureau, and regional health-care administrations. These institutions have been granted access by royal decrees based on Article 278 of the Programme Law of 24 December 2002. Access is intended for the performance of statutory tasks, such as policy preparation, monitoring and evaluation of healthcare. The Technical Committee of the Permanent Sample monitors the proper use of the data.

In principle, access for other organisations, such as universities or research institutions, is possible, but this requires a formal procedure:

- Application procedure: The organisation must apply stating the purpose and necessity of access;
- Assessment by the EPS Technical Committee: This committee evaluates the application based on criteria such as relevance, proportionality and data protection;
- Approval by the Sectoral Committee for Social Security and Health: If the assessment is positive, then the application is submitted to this committee, which oversees the protection of personal data;
- Royal Decree: If approved, then a royal decree is drawn up formalising access.

[16] The individuals concerned are followed up over the course of several years. However, there is an annual update of the panel: certain persons disappear from it (death, emigration) and new people with the same characteristics (gender, age) replace them; representativeness is, therefore, always guaranteed. Of all of the people up to 64 years old, 1/40 are in the sample and of all persons aged sixty-five or over, 1/20 are in the sample.

It is important to note that access to the EPS is subject to strict conditions in terms of both data security and privacy protection. The data is anonymised and it is not permitted to identify individuals.

The IMA does charge a commercial fee for access to data from the Permanent Health-care Sample. Access to the data is typically regulated and granted under strict conditions for non-commercial purposes, such as policy research and evaluation. Administrative or processing fees are possible, depending on the complexity and the effort required to prepare customized datasets.

3.5 Health Data Agency

Numerous data on health are collected: health care utilisation, mortality, causes of death, social security information, clinical data and health insurance reimbursement data are routinely collected for the management and operation of health and social security services. Information about the health status and health care of the population is fragmented across organisations and systems. Thus, there was a real need for an integrated federal initiative that facilitates the use and reuse of these data for health policy and health research.

The Belgian Health (Care) Data Agency was setup on the basis of the Act establishing and organising the Health (Care) Data Agency of March 14th, 2023, by five founding organisations (FAMHP, FPS, KCE, NIHDI and Sciensano).

The Belgian Health Data Agency's (HDA) mission is to facilitate access to quality health (care) data and data related to health (care) in a simplified and more uniform, reliable, transparent and secure manner, through the development of a framework in which the re-use (secondary use) of quality health (care) data and data related to health (care) is optimally facilitated; this is done in collaboration with regional and federal data holders and users. The HDA supports[17] the re-use of health and health-related data for population management, health prevention and healthcare policy, through research and innovation based on health needs and the active role of citizens, thereby contributing to better health and quality of life, within the framework of the tasks described in the Act of March 14th, 2023.

[17] To achieve this, the HDA focuses on the following objectives:

•Facilitate the availability of health (care) data and health (care) related data;

•Develop secure and reliable methods for the exchange of personal health data, with an emphasis on cybersecurity;

•Develop and implement a policy strategy on health (care) data and health (care) related data;

•Stimulate innovation, scientific research and policy support research that can contribute to better, affordable, preventive, and targeted healthcare;

•Ensuring the trust of citizens and patients by emphasising transparency and security around the re-use of health data, while respecting the rights and obligations of the parties involved.

HDA Data Catalogue: The Health Data Agency (HDA) aims to introduce a comprehensive data catalogue. This catalogue offers an overview of various datasets available from partner institutions. It also facilitates the general public in locating specific datasets and empowers the Health Data Agency to both assess and coordinate requests for data access requests. The HDA endeavours to make data Findable, Accessible, Interoperable and Reusable (FAIR). The HDA facilitates requests for data, both those that are documented in their catalogue as well as other relevant data requests. When a data request is submitted to the HDA, it will process the request.

The HDA provides detailed information about their activities and services in English on their website (including a Data request Form): https://www.hda.belgium.be/en

3.6 The PensionStat Platform

The PensionStat platform (PensionStat.be), is an initiative of the Belgian pension administrations, Sigedis,[18] the Federal Pension Service (FPS) and the National Institute for Social Security for the Self-employed (NISCE). They are supported by the Financial Services and Markets Authority (FSMA), the Federal Planning Bureau and by the FPS Social Security.

PensionStat provides aggregated, anonymous data on pensions in Belgium, based on the administrative databases of Belgian pension administrations, including from the Sigedis database on supplementary pensions (db2p) and the databases on statutory pensions of the Federal Pension Service (FPD) and the National Institute for the Social Insurance of the Self-employed (NISCE).

The platform provides up-to-date statistics that are easily accessible to a wide audience, from policymakers and researchers to pension experts and citizens. It covers not only statistics on statutory pensions, but also on supplementary pensions.[19]

[18] Sigedis is a non-profit public association that was founded in 2006 by social security institutions. This non-profit organisation fulfils a unique role as a one-stop data source, providing end-to-end solutions for anyone with data needs in the areas of person identification, care and retirement. More information (in Dutch or French): https://sigedis.be/fr/propos-de-nous/qui-sommes-nous

[19] These include figures about:

•the profile of Belgian pensioners (e.g., regarding the amount of their pension, their age, gender, pension scheme, pension type, etc);

•information on minimum pensions and guaranteed incomes for the elderly;

•information on supplementary pensions, such as number of members, participation rate, level of reserves, schemes, and types, and this for employees and self-employed persons;

•the inequality among recent retirees by gender, the so-called gender pension gap, for both statutory and supplementary pensions.

The administrations that took the initiative for PensionStat, supported by their scientific advisory committee, opted for a clear fixed structure per key figure with simple visualisations and objective explanations in accessible language. Each time these are accompanied by an explanatory glossary, transparent methodology and, importantly in this context, a reference to the underlying open data.

The PensionStat.be platform provides a comprehensive overview of statistical data on pensions in Belgium, but it is only available in French and Dutch. This includes information about both the data and methodology used on their platform, basis statistics and publications. Unless stated otherwise, the information published on the PensionStat.be website is free of copyright.

In order to access specialised pension data, such as detailed career information and information on second-pillar pensions, it is wise to contact Sigedis[20] directly. For information on public pensions and aggregated administrative data, it is wise contact the Federal Pension Service.[21] If your data request requires comprehensive or combined datasets from multiple administrative sources (such as integrated data from Sigedis and the Federal Pension Service for research purposes), then authorisation via the Crossroads Bank for Social Security (CBSS) may be required, particularly for extensive or sensitive data. In principle, no fee is charged, but a fee may be charged for costs related to the processing of the data request.

4 Conclusion

Belgium's commitment to leveraging administrative data for evidence-informed policymaking demonstrates the potential of harmonising data accessibility with robust privacy protections. The Belgian model underscores the importance of collaboration between regulatory authorities, researchers and policymakers, from the foundational frameworks established in the 1980s to the contemporary advances in data platforms such as Statbel and the Crossroads bank for Social Security.

The initiatives discussed in this chapter not only provide valuable insights for improving governance and public policy, but also serve as a reference point for other nations striving to balance innovation and accountability in the use of administrative data. Belgium's experience highlights the necessity of continuous refinement in legal frameworks and operational practices to sustain trust and societal benefits as the challenges of data privacy and technological advancements continue to evolve.

[20] Contact: https://sigedis.be/fr/contact

[21] Contact: https://www.sfpd.fgov.be/fr

Despite the pioneering role played by Belgium–certainly among continental European countries–in opening up administrative data for statistical and scientific purposes, there is currently a general feeling among institutional and scientific users of administrative data that both the regulatory and technical frameworks are no longer adapted to the enormous need for these data in the context of evidence-informed policy making and evaluation. The general perception is that access to administrative data has become more difficult and are available only after a lengthy procedure, which is due to a mixture of legal, financial and organisational constraints that cannot be traced exclusively to the introduction of the GDPR.

The ministers of the federal and regional governments responsible for labour decided to set up an interfederal consultation on the exchange of administrative data in order to optimise data exchange within and between policy levels, as well as with the scientific community (Interministerial Conference for Employment of December 1st, 2023). This consultation consists of a steering group of leading officials and a technical working group (BelAccess) of data experts from federal and regional administrations, and had the task of drawing up a vision report with concrete recommendations. On 19 June 2026, this report and the accompanying action plan were presented to the Interministerial Conference for Employment. Both were subsequently approved, and the Interministerial Conference for Employment mandated the BelAccess Steering Group to prepare an interfederal cooperation agreement on that basis.

Koen Vleminckx is a researcher and policy adviser with over three decades of experience in social protection and labour market policy. He began his career at the Herman Deleeck Centre for Social Policy (University of Antwerp) and later served as Operations Manager at the LIS Cross-National Data Center in Luxembourg, contributing to pioneering research on income inequality. After earning his PhD at KU Leuven, he held senior roles at Belgium's Federal Public Service Social Security, promoting evidence-informed policymaking. He co-authored influential publications on social investment and on child poverty, with the latter earning recognition and nomination for APSA's Aaron Wildavsky Award. He is currently Guest Professor of International Social Policy at KU Leuven and Secretary of Belgium's Interfederal Technical Working Group on Administrative Data Exchange, BelAccess, advancing access to high-quality data for policy innovation.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 4
On Some Aspects of Open Administrative Data in Bulgaria



Nikolay Netov

Abstract The primary sources of open data in Bulgaria include the National Statistical Institute (NSI), the official government portal for open administrative data (https://data.egov.bg/) and the Bulgarian National Bank (BNB). A considerable number of municipalities have initiated efforts to release their regional data, despite the lack of a cohesive national plan for regional open data sharing. Statistically aggregated datasets are also published routinely; however, there are commonly no explicit, transparent and universally accessible regulations governing access to microdata. An interesting contradiction in this context is the National Statistical Institute (NSI). Individual citizen initiatives cannot address the absence of open microdata from state and municipal entities.

1 Introduction

The provisions of Directive 2019/1024/EU have been integrated into Bulgarian law via the Access to Public Information Act and the Regulation concerning the standard conditions for the re-use of public sector information and its dissemination in an open format. Public sector organisations were mandated, by law, to develop and to progressively release, in an open format, a list of information arrays and resources that are freely accessible on the internet each year. The discussion surrounding the re-use of public sector information, in an open machine-readable format, is elaborated upon in both the e-Government Act and in the Regulation on the general requirements for information systems, registers and electronic administrative services.

The legal framework for access to statistical microdata in Bulgaria is established by the Statistics Act, which stipulates in Article 25, paragraph 1, that individual data is deemed statistical secret and may be utilised solely for statistical purposes. The Regulations governing the provision of anonymised microdata for scientific and research purposes, accessible on the National Statistical Institute (NSI) website (https://www.nsi.bg/en/documents/detail/52), define the terms and conditions for supplying

N. Netov (✉)
University of Sofia, Sofia, Bulgaria
e-mail: nnetoff@feb.uni-sofia.bg

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_4

anonymised microdata to individuals who are engaged in scientific research. This process is administered by the NSI exclusively, following strict security protocols and safeguards against re-identification.

It is crucial to recognise that specific procedures are enacted in order to prevent the utilisation of shared unique identifiers across sources and the potential to integrate disparate data sets both within and among various data-holding entities.

With the exception of the National Statistical Institute, the other Bulgarian state institutions (1) do not provide clear information about the microdata available to them; and (2) do not have established procedures concerning their provision. In this regard:

1. Researchers can find out where and what administrative data are available on Bulgarian state institutions through the RegiX information portal. All available electronic datasets, as well as information about the metadata associated with them, can be found, classified by Bulgarian state institutions, at https://info-regix.egov.bg/public. Officially, this information is available only for data exchange between individual government institutions; however, in practice it is also a complete catalogue of the information that researchers can request;
2. Generally, in Bulgaria, every interested microdata user must contact the relevant microdata-holding organisation in order to obtain access, with each case being addressed on an individual basis and in a fashion devoid of overarching regulations. Each state institution's official website offers an email address for correspondence, where you can submit a request for access to the required microdata. Generally, you will obtain an invitation to visit the relevant institution in person and to articulate your desire to an assigned representative. The practices for granting or denying access are case-based.

Following the Introduction, in Sect. 2 we delineate the Bulgarian government's official open administrative data portal. Section 3 presents an analysis of the policies and practices regarding access to open datasets and microdata provided by the Bulgarian National Statistical Institute. Section 4 elucidates the data sets that are accessible through Bulgarian National Bank's website. Section 5 presents other public initiatives, while Sect. 6 presents individual initiatives in the field of open microdata. The conclusions are outlined in the final section.

2 Official Government Open Administrative Data Portal

The official government open administrative data portal can be found at https://data.egov.bg/. It is published in Bulgarian. The English version is translated only partially. At the time of writing, 11,678 datasets from 543 state institutions have been published. In general, these data sets are snapshots of certain registries at a certain moment. One can often find a time series for specified time periods. There are about 17,000 registries in Bulgaria, but most of them are on paper and this makes the publication of open administrative data uneasy. Municipalities maintain

similar registries, but decisions about which to open are inconsistent. Unfortunately, data sets are organised in a somewhat chaotic fashion and it is not easy to find a specific data set. Many data sets are published by municipalities, but they are listed randomly. Different municipalities publish data from different registries according to local decisions. Still, the most important registries are digital, and some data are available as open administrative data. The data sets are organised into 14 categories, but 3826 data sets have not yet been categorised. The main areas covered are:

- Agriculture, fisheries and aquaculture, forestry and food;
- Education, culture and sport;
- Environment;
- Energy;
- Transport;
- Science and technology;
- Economy and finance;
- Population and social conditions;
- Government and public sector;
- Health;
- Regions and cities;
- Justice, legal system and public safety;
- International affairs;
- High-Value Datasets (HVD).

Most of the data sets are published in .csv format (7642), json (569) and xml (117). There are also zipped data sets (31). There are options to transform and download a data set in any of the formats mentioned above. For example, the Commercial Registry publishes their incremental data day-by-day in xml format. Visualisations in browser are also available for some data. There is no consistency and other data sets are available only in the machine-readable formats.

The government's open administrative data portal strictly adheres to the anonymisation protocols for published datasets, as mandated by the legal framework. Determining the proportion of public microdata within predominantly published statistically aggregated datasets is also quite challenging.

3 National Statistical Institute (NSI)

The National Statistical Institute (NSI) is another source of data (https://nsi.bg/en). The National Statistical Institute's information System INFOSTAT provides statistical information on the Republic of Bulgaria's demographic and socio-economic conditions and development. The system provides users with the possibility to create tables with data on various indicators, as well as to visualise the results graphically. A browser is needed to access data and visualisations. A Business Intelligence (BI) functionality is built into the website. The system allows for the results and their metadata to be downloaded in the following formats: xlsx, csv and in pdf. Data and

visualisations are much more consistent and up to date than those published via the official government's open administrative data portal. The NSI's site is bilingual and all publications are accessible in both languages. Switching the interface from one language to another is instant and data is not affected by the change.

NSI provides statistical data in the following areas:

- Business Statistics;
- Demographic and Social Statistics;
- Macroeconomic Statistics;
- Regional statistics and indicators for monitoring;
- Environment and Energy;
- Agriculture and Forestry;
- Tax statistics;
- Key Indicators.

A methodology for gathering data and metadata is available for every published data set. Users are provided with paid access to more detailed data at lower levels, after disaggregation, and to custom-tailored data sets.

The NSI has established publicly accessible regulations for the provision of micro-data.[1] In order to obtain a particular set of microdata, please direct your request to the email address info@NSI.bg. A representative from NSI will reach out to you and request documentation to verify your eligibility as a microdata user. The following users can apply for access to the anonymised microdata:

- universities and other higher education institutions;
- organizations and scientific research institutions;
- other agencies, organizations or institutions that have specific structural units that carry out scientific and research activities;
- other legal entities that use data for research and analytical purposes;
- natural people that need anonymised data for the purposes of scientific and research activities.

The specialised expert council will make the final decision about your application, reviewing it within a timeframe that does not exceed one month. A price quote will be sent to you following the application's approval. Upon acceptance of this quote and subsequent payment, NIS will proceed to prepare a data set request on your behalf. In order to access your data set, you must first sign a sworn declaration affirming that the data will be stored and utilised in compliance with the established rules. Receiving the data on site is not a requirement, and you may even be provided with remote access to it. The data is encrypted and the information necessary for decryption is supplied securely.

The microdata sets accessible from the NSI are limited to the datasets available in their information system INFOSTAT. The NSI's INFOSTAT information system provides statistical information about the demographic and socio-economic status and development of Bulgaria. Users can create tables with data on various indicators

[1] https://www.nsi.bg/en/documents/detail/52

and can visualise them graphically. The system allows the results and their metadata to be downloaded in the following formats: xlsx, csv and pdf. The list of standard statistical indicators distributed by the NSI is organised as described in Box 4.1:

Box 4.1: List of statistical indicators distributed by the National Statistical Institute Business statistics: Standardized business surveys, business register, foreign trade, annual business statistics, investments, short-term business statistics, R&D, innovations and information society, tourism, transport

Demographic and social statistics: Households income, expenditure and consumption, housing fund, health, culture, population (demography, migration and projections), education and lifelong learning, labour market, justice and crime, census, social protection, social inclusion and living conditions

Macroeconomic statistics: Gross Domestic Product (GDP), Gross National Income (GNI), Government Finance Statistics, Employment and Hours worked—total of the economy, Employment and Hours worked—by Regions, Inflation and Consumer Price Indices, National accounts—quality reports for the ESA 2010 Data Transmission Programme, Purchasing Power Parities (PPPs)—European and International Comparisons Programme, Pension entitlements in social insurance, Non-financial national accounts by institutional sectors, Housing price statistics, Supply Use Tables, Financial national accounts

Energy: Production and deliveries of energy products, electricity and natural gas prices, combined heat and power units (CHP units), energy balances—energy supply block, overall energy balance sheet, disaggregation of final energy consumption, district heating and cooling systems, electricity and natural gas market indicators, classification PRODENERGY

Environment: Air emissions, air emissions accounts, Waste from economic activity, municipal waste, registered noise levels, protected natural scenery, expenditure on protection and restoration of the environment, tangible fixed assets with ecological use, water statistics, Total revenue from taxes and fees and expenditure for municipal waste, packages placed on the market, environmental taxes, underground reserves, environmental goods and services, physical energy flow accounts

Multi-domain statistics: Crisis Events, Indicators for Europe 2020 Strategy, Regional Statistics, Land use distribution of the Republic of Bulgaria, City Statistics, Sustainable Development

Georeferenced data: Population grid

Agriculture: Prices and producer price indices in agriculture—National level, prices indices of goods and services currently consumed in agriculture and goods and services contributing to agricultural investment, agricultural land market and rent, economic accounts for agriculture—Annual data

Forestry: Producer price indices in forestry, producer prices in forestry

Sales of liquid fuels: Petrol, diesel and liquefied petroleum gas

4 Bulgarian National Bank

The Bulgarian National Bank[2] is another big supplier of open administrative data. Most of the data sets are published as either Excel files or in CSV format. At the time of writing, 14 registers are available on the BNB's Website, three of which are in Word format, and one of which is in HTML format.

Six data sets are available at the section 'Statistics on banknotes and coins in circulation.' It contains monthly information, starting from December 1999, about the dynamics and structure of banknotes and coins in circulation, as well as about quarterly data on the number of non-genuine banknotes retained and about the circulation coins starting from the first quarter of 2007. The data are organised in a time series.

The section 'Statistics' provides access to many data sets of macroeconomic and financial data. Most of the datasets are organised in a time series.

Some of the data published on the BNB's Website are presented in the form of reports and are in pdf format. The bank itself uses these datasets to conduct research and to create reports. The government, foundations, financial institutions, and think-tank organisations use the published data sets in their day-to-day operations, analysis, and development of strategies. The banking sector, financial and non-financial institutions heavily rely on publicly available statistical and financial data to make marketing and financial decisions (e.g., interest rates, etc.).

The most interesting, from the perspective of microdata users, is the Central Credit Register.[3] The establishment and functioning of the Central Credit Register is regulated by Regulation No. 22 on the Central Credit Register (CCR). The Register is an information system that is organised and maintained by the Bulgarian National Bank (BNB) for clients' credit indebtedness to banks, financial institutions, payment institutions and electronic money companies, granting loans under the procedure of Art. 21 of the Payment Services and Payment Systems Act (PSPA), which operate on the territory of the Republic of Bulgaria, as well as to investors who have provided a loan under a project through a crowdfunding service provider; there is an exception for foreign providers who operate directly on the territory of the Republic of Bulgaria. The Register assures:

- The centralisation of information about the credit indebtedness of clients to banks, financial institutions, payment institutions and electronic money companies, as well as to investors who have provided a loan under a project through a crowdfunding service provider;
- use of information from banks, financial institutions, payment institutions and electronic money companies, as well as from crowdfunding service providers on the credit indebtedness of their clients;
- summarising the collected information and using it for the needs of the BNB.

[2] https://www.bnb.bg/index.htm

[3] https://www.bnb.bg/RegistersAndServices/RSCreditRegister/index.htm

The Central Credit Register maintains information about people who are co-debtors and guarantors of loans.

The register records loans made to clients from banks and financial institutions, to payment institutions and electronic money companies, as well as to investors who have provided a loan under a project through a crowdfunding service provider, regardless of their size.

Banks and financial institutions, payment institutions and electronic money companies, as well as crowdfunding service providers, submit information on the status of all the active loans of their clients as of the last date of the month to the Central Credit Register monthly, by the 15th day of the month following the reporting month. Banks, financial institutions, payment institutions and electronic money companies, as well as crowdfunding service providers, also submit information about adjustments made to their clients' loans for previous reporting periods within the same deadline.

5 Other Public Initiatives

5.1 Ministry of Finance

As a result of extremely strong public pressure, the Ministry of Finance has started providing open microdata for the main state and municipal payments through the Electronic Budget Payments System (SEBRA).[4] SEBRA is a system for the monitoring of payments initiated by legal entities included therein and for managing payments within pre-set limits. The SEBRA information does not include:

- Payments from foreign currency accounts of budgetary organizations and accounts opened for servicing funds under international programmes and agreements, in the presence of an explicit requirement for a separate bank account in the relevant EU treaties and regulations;
- Payments made under the scheme for centralised payment of social security contributions pursuant to Art. 159 of the Public Finance Act;
- Operations that are subject to inclusion in the reports for the implementation of the relevant budget or the accounts for funds from the European Union, which are equated to cash flow in accordance with the requirements for the reporting of budgetary organizations on a cash basis pursuant to item 7.4 of VAT Act 20/2004 (financial leasing, etc.).

The operations indicated above represent a resource of an insignificant volume.

[4] https://www.minfin.bg/bg/transparency/2025-06-12

5.2 Ministry of Environment and Water

As a result of EU funded projects, the Ministry of Environment and Water has started providing open microdata from Bulgaria National Air Quality Monitoring System.[5]

The system for providing the public with information about ambient air quality in real time serves to present data collected and processed through the National Automated Air Quality (AQ) monitoring system. The system serves to present it in an easily understandable format, due to the specialised nature and public importance of this type of information. For the purpose of providing information on AQ, the Information System is based on the norms set out in national and European legislation, as well as an approved European-level AQ index published on the European Air Quality Index (European Air Quality Index—http://airindex.eea.europa.eu/). Following the trends imposed by the European Air Quality Index, the Information System for Providing the Public with AQI Data uses geographic maps to visualise the measurement points of the National Automated AQ Monitoring System, as well as the ability to view the concentrations of the relevant pollutants, measured at a given point. The data about the concentrations measured are visualised using graphs. The information system provides, in addition to the possibility of viewing data in real time, the possibility of preparing reference information for a period regarding the hourly, daily and annual average values measured for the relevant pollutants. The visualisation of information on atmospheric air quality is carried out based on real-time data from the National Automated System for Monitoring of Air Quality. These data are preliminary and are subject to verification. The final data is published in the quarterly and annual bulletins that are issued by the Executive Environmental Agency (EEA).

5.3 Ministry of Transport (MT)

The most interesting, from the perspective of microdata users, is the MT register of vehicle technical inspections and MT register of the public transport timetables. Unfortunately, access to this data is only possible with an approved request under the Access to Public Information Act[6] (APIA). Additional information regarding this procedure can be obtained by visiting the following link: https://www.aip-bg.org/en/howto/faq/.

5.4 Ministry of Health

The Ministry of Health publishes health statistics, hospital information and epidemiological reports. Access to microdata is uncommon and is generally restricted.

[5] https://eea.government.bg/kav/

[6] https://www.aip-bg.org/en/legislation/Text_of_the_APIA/200432/

5.5 *Ministry of Labour and Social Policy*

The Ministry of Labour and Social Policy publishes data on employment and social assistance. Microdata from surveys might be made available through collaboration with the NSI.

5.6 *Ministry of Education and Science*

The Ministry of Education and Science provides statistics and performance data at the school level. Microdata, such as student-level anonymised data, is not publicly available; however, it can be accessed upon request for research purposes.

5.7 *Local Authorities*

Local authorities participate in the Open Government Partnership (OGP) initiative.[7] Most of them aim to publish more machine-readable datasets (e.g., transport, urban planning and public services). Microdata is generally not published, but may be made available with an approved request made under the Access to Public Information Act (APIA).

6 Individuals' Initiatives

In Bulgaria, one of the proponents of open administrative data is Boyan Yorukov.[8] He is well-known for making visualisations of publicly available data. What is much more important, though, is that he also gathers data from different sources, transforms and publishes combined data sets and visualisations. For example, recently (2024), Boyan succeeded in collecting data from government, municipality and utility companies for about 322,000 shafts in Sofia. These underground shafts were put on a map as is visible in Fig. 4.1.

The more important point is that the map is interactive, and one can see the building permit with which a particular shaft was built. One of the reasons for this activity is that the shafts are badly maintained and built chaotically, often without synchronisation among the organisations involved. Thus, these create a lot of trouble for citizens and cost large amounts of money. This work has inspired a lot of social activists and pressures the municipality and utility companies to take action to improve the situation.

[7] https://www.opengovpartnership.org/

[8] https://yurukov.net/

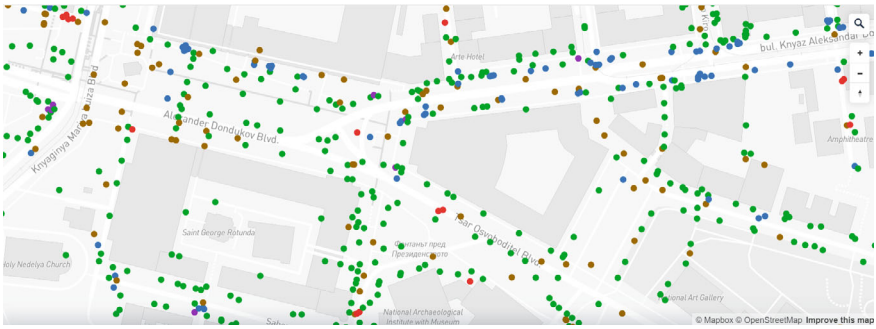


Fig. 4.1 Visualization of underground shafts in the center of Sofia

Many other examples of visualisations are available at https://opendata.yurukov.net/. It is important to mention that Boyan Yurukov himself generates open administrative data processing data from publicly available documents that are not in machine-readable formats. The resulting data sets are available in machine-readable format and are available for free.

7 Conclusion

Bulgarian governmental agencies consistently provide a substantial volume of statistically aggregated open datasets. Regrettably, we are unable to reach the same conclusion concerning microdata. There are commonly no explicit, transparent and readily accessible regulations concerning the access to microdata. An important exception in this context is the National Statistical Institute (NSI). In conclusion, we must underscore that individuals' initiatives in this regard cannot rectify the deficiency of accessible microdata from governmental and local entities.

The obstacles confronting the Bulgarian open microdata ecosystem are complex. Factors include the legislative framework, bureaucratic inertia, poor understanding among government servants on the advantages of open data and inadequate training in open data management impede advancement.

Nikolay Netov is Head of the Department of Statistics and Econometrics at the Faculty of Economics and Business Administration of Sofia University. Over the last fifteen years he has worked in a number of international scientific, applied and educational projects in collaboration with a large number of colleagues from all over the world. In 2014 he received the annual award for the project "Career in IT", initiated and led by him, as the best project for 2013, implementing cooperation between business and NGOs.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 5
The Reality of Administrative Data in Croatia



Nataša Nater Drvenkar, Alka Obadić, and Josip Tica

Abstract This chapter discusses existent administrative data collection methods throughout the Republic of Croatia (hereafter Croatia). Data collection, processing and publication in national rules are aligned with EU guidelines. State and public bodies allow free access to statistical data, while selected groups (researchers and scientists) can also obtain access through specific requests and others can gain access via the Croatian Bureau of Statistics for a fee. Data collection, analysis and publication frequency depend upon both the data type and the collecting body. While published data is mostly documented in English, microdata access is exclusively in the national language. The key challenge arises with microdata, an arena in which there is still much room for improvement.

1 Introduction

This chapter outlines data collection methods in Croatia and focuses on two main areas: (1) the collection, analysis, availability and relevance of administrative sources, and (2) specific case studies that are important for academic and professional research. Since joining the EU, many state and public institutions (Croatian Bureau of Statistics (CBS), Croatian Employment Service (CES) and the Croatian National Bank (CNB), etc.) have provided centralised online access to their data through the State Agency for the Development of the Digital Society. It is expected that this agency's internet service will become the primary access point for all administrative data in Croatia. Currently, no single shared admin data access point exists in Croatia,

N. Nater Drvenkar (✉)
Faculty of Economics and Business, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia
e-mail: natasa.nater@efos.hr

A. Obadić · J. Tica
Faculty of Economics and Business, University of Zagreb, Zagreb, Croatia
e-mail: aobadic@efzg.hr

J. Tica
e-mail: jtica@efzg.hr

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_5

greatly limiting the availability of microdata to interested researchers and policy-makers alike. This fragmented system means that accessing these valuable datasets is often both a challenging and time-consuming process. Generally, access to microdata is primarily granted to domestic scientific institutes and universities, which typically have the credentials and resources necessary to navigate the complex administrative requirements involved. Additionally, international research projects that collaborate with these local institutions may also secure access, but often under specific conditions and through more extensive agreements. As a result, many potential users, who might benefit from this data, such as independent researchers or smaller organisations, face significant barriers to obtaining the necessary permissions. Future access will depend on the development of EU regulations, such as the Data Act and related laws, as well as on national choices.

In Croatia, individual official institutions manage data relevant to their respective areas and use privileged access to administrative data from other state bodies in order to collect, process and publish it. Much of this data is publicly available online, while some can be accessed only by authorised users via web services. The State Agency for the Development of the Digital Society is working to create a centralised data access point that will become the main access point for all administrative data. Legal and practical restrictions limit the release of data collected by statistical authorities, such as the Croatian Bureau of Statistics or the Tax Administration. Data requests must usually be addressed to the original owners, especially when confidential, personal or sensitive data is involved. However, non-confidential data can be requested by any natural or legal person under the Right of Access to Information Act through various channels, including websites, publications, databases, emails or specific requests to the competent authorities.

Administrative data have profoundly influenced social science research and have proven the crucial role they play in improving both research methods and outcomes, despite challenges related to privacy and data protection (Connelly et al., 2016). This data is collected by government agencies for purposes such as tax collection, social security, education, employment and health and is essential for both official, administrative, policy and scientific research purposes as well as for the design and evaluation of political measures (Crato & Paruolo, 2019).

Data security includes the protection of employee data compliance with regulations such as the GDPR and the introduction of technological and institutional measures and is ensured. This integrated approach emphasises secure data processing as an essential prerequisite for effective and equitable labour market policies. Ritchie (2008) emphasised the importance of administrative data for decision making, given that it provides important insights for the analysis of trends in the health, economic and social sectors. He highlighted the Virtual Microdata Laboratory (VML) set up by the Office for National Statistics in the UK, which provides academic researchers with secure access to sensitive data, thereby promoting policy-relevant research. He also pointed out the challenges associated with sensitive microdata, which includes personal and commercial information where there is a risk of identification (Müller & Umkehrer, 2023).

Several researchers, including Duncan and Pearson (1991), Dale (2005), Crato (2017), and Corrado (2019), identified key challenges in dealing with microdata: (1) the risk of data disclosure, (2) balancing security and research efficiency, (3) using data for non-statistical purposes, (4) limitations on the details of datasets and geographic information and (5) maintaining secure data access settings. The former can be identified in terms of the problem of microdata collection in Croatia. For example, basic challenges in collecting such data include (1) sensitivity, invasion of privacy, compliance with laws (e.g. GDPR), ethics, trust and voluntary consent, (2) data bias, methodological errors, lack of knowledge and lengthy approval procedures, (3) heterogeneous samples, (4) the availability of resources for data collection, processing, analysis and dissemination (experts, finance, technology and cybersecurity) or (5) linking micro-level data from different sources and ensuring temporal consistency. Ritchie's example of the VML (Fuchs et al., 2023) in the UK and the Austrian "AMDC" research ecosystem (see Chap. 2) show effective models that successfully balance access to detailed data with the protection of sensitive information. The remainder of this chapter explains the development of the Croatian administrative and microdata system principles.

2 Development and Characteristics of the Administrative and Microdata System in Croatia

In Croatia, the use of administrative data for research and policy-making has changed from a centralised system, during the Yugoslav era, to a more open and integrated approach. This change, influenced by the demand for evidence-based decision-making, technological advances and alignment with international standards, has improved the transparency, availability and quality of data. Improvements in microdata access policies and practices, such as better metadata standards, advanced anonymisation techniques, and expanded access are necessary to meeting the evolving needs of policy-makers and researchers. The main official institutions responsible for data collection, processing, analysis and availability include the Croatian Bureau of Statistics, Croatian National Bank, Croatian Financial Agency (FINA), Ministry of Economy and Sustainable Development, Croatian Institute of Public Health, Ministry of Finance and Ministry of Agriculture, as well as other statistical authorities. Detailed historical and legal frameworks are presented in a timeline (Table 5.1). This timetable reflects Croatia's commitment to improving access to microdata for policy-relevant analysis, thereby ensuring that statistics used in policy are reliable, accurate and accessible while maintaining the necessary protection of individual privacy.

Table 5.1 Legal framework and EU harmonization of Croatian administrative data

Legal framework/ institution foundation	Name of legal framework/ institution foundation	Key areas of activity and Microdata type
1992	Croatian Bureau of Statistics (CBS)	The CBS was established to collect, process and to disseminate statistical data as they are the main institution responsible for access to official statistics in Croatia; "on site" mode and databases are in the transitional (research) phase of availability
1993	Establishment of the Croatian financial agency ("Fina")	A specialised agency for the provision of financial services, payment services, management of accounts and settlement of financial transactions, development of online platforms for business users and maintenance of a register of financial statements of legal entities
2003	The Development Strategy of Official Statistics of the Republic of Croatia 2004–2012	Croatian statistical system has been significantly improved in the preceding period and in the period of negotiations for Croatian membership in the EU; it has also been harmonised with EU legislation and standards
2009	The official statistics act (official Gazette, no. 25/20; regulation 223/2009 of the European Parliament)	The basis for the functioning of the CBS and the statistical system in Croatia. It contains provisions on confidentiality, protection and access to microdata and is adopted for a period identical to the period defined in the European statistical Programme
2010	The code of practice for European statistics	Croatia has adopted the European statistics code of practice, which emphasises the importance of professional independence, impartiality and transparency in statistical work. This adoption harmonises Croatian practices with EU standards, including access to microdata
2013	Accession to the European Union	Croatia's accession to the EU brought about significant changes in statistical practice to ensure compliance with EU regulations and standards. This step facilitated the harmonisation of procedures for data collection, processing and dissemination, including the availability of microdata

(continued)

Table 5.1 (continued)

Legal framework/ institution foundation	Name of legal framework/ institution foundation	Key areas of activity and Microdata type
	The Development Strategy of Official Statistics of the Republic of Croatia 2013–2022 (Official Gazette 69/13)	
	The act on the right of access to information (official Gazette, no. 25/13, update: 85/15, 69/22)	This document outlines the importance of the protection of rights and the circumstances of data availability (e.g., the right to copy, publish, disseminate and communicate information to the public, to process information, to use information for commercial and non-commercial purposes, and the obligation of users to indicate the source of data)
2014	ESS European statistical system	Authorities responsible for the development, production and dissemination of European statistics in the EU27 and important for the harmonisation of the methodological framework and dissemination of the impact of the link; access mode "on site," but files can also be sent to authorised researchers and can be used on the premises of research institutions; the databases are in the advanced (research) phase of availability
2017	Revision of the official statistics act	The revised law on official statistics contains clearer provisions about access to microdata for both scientific and research purposes. The aim was to strike a balance between data protection and the need for data availability for policy analysis
2019	The Croatian Centre for the Development of access to microdata	Establishment of the Croatian Centre for the Development of access to microdata to support researchers and policy makers by facilitating access to anonymised microdata sets
2020–2021	Adoption of the National Programme for the Development of Official Statistics 2021–2027	This programme outlines strategic goals for the development and improvement of the statistical system, including further improvements in terms of the availability of microdata

(continued)

Table 5.1 (continued)

Legal framework/ institution foundation	Name of legal framework/ institution foundation	Key areas of activity and Microdata type
2022	Development Strategy of Official Statistics of the Republic of Croatia 2021–2030	It is based on the relevant regulations of the Republic of Croatia, the European Union and international agreements in the field of official statistics in order to increase the availability of statistical data, their quality, timeliness and comparability
2023	Annual implementation plan of statistical activities of the Republic of Croatia for 2023	This short-strategic planning act is important for further coordination and cohesion and implementation of ESSM

Note The collection and processing of data prior to Croatian independence was influenced by the political, economic and social circumstances in Yugoslavia at that time. The main characteristics were: (1) a centralised statistical system with republican statistical institutes working according to federal guidelines, and (2) data collection was based on a planned economy with a strong emphasis on economic data such as production, consumption, employment and social activities
Source Compiled by the authors based on research into the Croatian legal framework from 1992 to 2024.

3 Important Institutions for the Collection, Processing, Analysis and Availability of Aggregated Data and Microdata in Croatia

3.1 Croatian Bureau of Statistics (CBS)

The Croatian Bureau of Statistics (CBS) is the most important institution for the production, dissemination and coordination of official statistics in Croatia and represents the national statistical system in European and international bodies. The CBS operates autonomously, based on the Official Statistics Act, and its activities are governed by the Programme for Statistical Activities and the Annual Implementation Plan. The Croatian statistical system provides a wide range of internationally comparable socio-economic data and follows global best practices and focuses on efficiency, user orientation and human resources development. In addition to the CBS and the Croatian National Bank, the Ministry of Economy and Sustainable Development, Croatian Institute of Public Health and the Ministry of Finance and Ministry of Agriculture are also official producers of statistical. The programme for statistical activities is aligned with the European Statistical Programme and the CBS uses different approaches to access administrative data to reduce the reporting burden and to speed up statistical processes.

All published data is available on the website,[1] which is bilingual (Croatian and English). There are six publication options: (1) First publications, (2) Statistical reports, (3) Statistics in line, (4) PC-Axis databases, (5) STS database and (6) Geostat. All publicly published data on the CBS website is aggregated. The Croatian Bureau of Statistics (CBS) offers special data processing on request, which may be subject to a fee. Older publications, which are only available in printed form, can be consulted in the CBS library. Data are published according to different schedules—daily, weekly, monthly or thematically—with publication dates being defined and detailed in the 2024 Data Publication Calendar and in the 2024 Publication Programme. These schedules and programmes are published annually in Croatian and English on the CBS's website free of charge. The publication programme outlines the types of publications available and provides a detailed schedule for the entire year and into the following May, depending on the completion of the statistical surveys. More detailed microdata is not publicly available. Public use microdata files contain information about individual statistical units and are prepared in a way that prevents the identification, either directly or indirectly, of the statistical unit, thereby considering all relevant means by which a third party might potentially identify it.

Access to confidential statistical data of the CBS (which also consider micro (administrative) data) for scientific purposes can only be obtained by research institutions–legal entities carrying out scientific research activity or scientific research units within a legal entity, which must be independent and autonomous in the formulation of scientific conclusions and separate from the field of activity of the legal entity to which they belong. Applications for access to confidential statistical data for scientific purposes are considered by the Statistical Confidentiality Committee of the State Statistical Office, in accordance with the Conditions and Methodology for Access to and Use of Confidential Statistical Data for Scientific Purposes- Basic Information. Currently, the CBS does not provide for the possibility of combining different individual sources into "one larger one," (i.e., that the CBS has unified all data from different institutions according to the same methodology and is then a kind of "service" for providing this information (in the sense that the researcher does not then have to search for the institution and the data from its jurisdiction individually)). Access to data on portable storage media enables access to the databases of the following statistical surveys: Labour Force Survey (LFS) Income Survey (IS) Household Consumption Survey (HCS) Innovation Activities of Enterprises (IAE) Annual Survey on the Use of Information and Communication Technologies in Households and by Individuals (ICT-HH). The Central Bureau of Statistics decides on the manner through which to access confidential statistical data for scientific purposes, taking the level of detail and the 'sensitivity' of the statistical data for which access is requested into account. The whole CBS website is bilingual (Croatian and English versions) and everything is available in both languages. The CBS develops, collects, produces and disseminates official statistical data in accordance with the applicable national and European legal framework and can be consulted in detail. It does not combine

[1] For further information, see: https://podaci.dzs.hr/en/.

the various data sources into a unified data set, so researchers must collect data from different institutions individually.

3.2 Croatian Employment Service

The Croatian Employment Service (CES[2]) is a public institution owned by the Republic of Croatia, established by the Labour Market Act (Official Gazette 118/2018; 32/20; 18/22), which plays a key role in regulating supply and demand in Croatia's labour market. The main task and purpose of the CES is to effectively mediate in the labour market in order to provide the necessary workforce, achieve full employment and exercise rights during unemployment, while promoting partnerships with labour market actors. The CES operates through its central office, five regional centres (Osijek, Rijeka, Split, Varaždin and Zagreb), 18 regional offices and 99 local offices and provides its services throughout the territory of the Republic of Croatia (CES, 2024a).

The Ministry of Labour, Pension System, Family and Social Policy oversees the legal activities of the Croatian Employment Service (CES). In May 2022, the CES underwent a comprehensive restructuring and adopted new internal organisational rules. Significant changes were made in both analytics and informatics to support the digitalisation and the modernisation of CES business processes. Analytical and statistical functions were transferred from the IT department to a newly established, independent department for analytics, statistics, research and business reporting and this department now reports directly to the CES Director. Two new departments were also created as part of the reorganisation: the Personal Data Protection and Information Security Department and the IT Management and Infrastructure Monitoring Department (HZZ, 2023, p. 51). The Croatian Employment Service (CES) also focuses primarily on collecting data about labour supply, on unemployment and about job seekers. The CES records registered unemployment and makes this data available daily on its website in Croatian and English under "Labour market statistics—open data." The CES also collects data on persons newly added to the register and employees who have left the register. New entrants are unemployed people that registered with the CES in the month under review. Employed people from the CES register are people who were employed during the reference month under an employment contract or through other business activities and who were previously registered with the CES.

The Croatian Employment Service (CES) is the only institution in Croatia that is responsible for collecting data on the users of active and passive employment policy measures. Passive employment policy measures refer to recipients of unemployment benefits granted to unemployed persons registered with the CES who meet the criteria set out in the Employment Services and Unemployment Rights Act (Official Gazette, 2018). Active Labour Market Policy (ALMP) measures, which are regulated by the

[2] In Croatian: Hrvatski zavodza zapošljavanje—HZZ.

Labour Market Law (Official Gazette, 118/18, 32/20), include promoting employment, providing training, supporting business start-ups, improving education and offering on-the-job training (see also Obadić, 2016; Obadić, 2020). These measures also focus on job creation through public works and support for people at risk of losing their jobs. Data on people participating in ALMP measures are collected monthly in 21 districts, with variables such as new entrants, leavers and active participants recorded at the end of each month. The statistics are collected on an ongoing basis and provide a comprehensive record of active participants from both the previous year and the current year (CES, 2024b).

The Croatian Employment Service (CES) also collects comprehensive data on labour demand. This data comes directly from employers who report vacancies for various types of contracts and economic sectors to the CES. These vacancies are recorded monthly and provide information about both the distribution of employment opportunities by location both within the country and internationally. The CES database processes this data efficiently, culminating in the monthly publication of key statistics in the Monthly Statistical Bulletin. In addition, a quarterly Analytical Bulletin takes a closer look at market developments by analysing the impact of macroeconomic indicators on specific themes. The CES also publishes a Yearbook that focuses on statistical data from the previous year. In addition to these publications, the CES also offers an interactive online statistical database through which users can access data on unemployment according to their preferences. This database is currently only available in Croatian, but an English version is planned. For users who are unable to find certain data via the online platform, the CES responds promptly and free of charge to requests by email in accordance with the Law on the Right of Access to Information. The CES attaches great importance to data accessibility, maintaining transparency and protecting personal and sensitive information. Cooperation with institutions, electronic data exchange and the obligation to cooperate all further facilitate the exchange of information. The CES's website provides general information and selected statistics in English, thereby ensuring the dissemination of information to a wider audience while protecting data privacy by restricting access to personal data and microdata.

4 Challenging Areas of Administrative Sources and Microdata System in Croatia

In Croatia, various professional associations use methods and tools to effectively analyse sector-specific data. Institutes such as the Institute for Tourism, the Institute for Public Finance and the Institute for Development and International Relations all conduct readiness analyses and statistical reviews in their respective fields. The Croatian Employers' Association focuses on research into digital transformation. One example of this is the "Croatian Digital Index" study by Apsolon, which assesses the economy's readiness for digital progress. Organisations such as the Croatian

Pension Insurance Institute and the Croatian Tax Administration provide accessible online resources, which comply with EU web accessibility directives, and provide information about insured persons, pensions, taxes and international tax regulations. The Croatian Chamber of Commerce, as the largest business network in Croatia, supports its members with services such as business networking, supplier databases and business information to improve the competitiveness of companies. The Croatian Employers' Association has also represented the business sectors since 1994 and participates in the tripartite social dialogue for the protection of economic and social rights through the Economic and Social Council. It uses official data from public institutions to carry out customised sector analyses for its members. However, much like at the national level, the tracking of regional micro-data and its collection, processing and publication is particularly pronounced. The Ministry of Regional Development and the EU funds provide information on the Development Index, most of which is available in Croatian. Therefore, access to microdata in Croatia varies according to risk assessment, research needs and data availability, as shown in Table 5.2. While most data are publicly available, certain information may require customised analysis or through a request being made to the respective agencies.

As mentioned above, some of the key points about the different approaches to microdata access in Croatia are the challenges associated with working with microdata, such as the risk of disclosing data, balancing security measures with research opportunities, concerns about the use of non-statistical data, limitations on details and geographical information and the importance of maintaining secure settings for data access. By paying attention to these specificities, EU countries and Croatia

Table 5.2 Different approaches to microdata access in Croatia

Types of data		Challenges
Official and original microdata (CBS and so on)	Access to databases Access to big data Available microdata with verification by researchers	Lack of data harmonisation Lack of "open" availability of microdata Coordination costs for finances, experts, but also time Different methods for the collection, recording, processing and analysing data (e.g. labour market, legal entities, GDP at national and regional—NUTS level)
Special surveys	Limited individual access	Scope of data collection Methodology Sampling problem (heterogeneity/homogeneity) Costs
Commercial microdata	Direct individual (but closed) access	Scope of data collection Methodology Sector specificity and individual data (financial analysis) Costs

Source Compiled by the authors

are striving to produce reliable, detailed and high-quality microdata that are invaluable for both research and policy-making. Researchers, particularly economists in Croatia, often base their analyses on various types of microdata. However, even in this area, most data are based on administrative aggregate data. In addition to the outstanding problem of collecting, analysing and providing microdata, the following presents some specific areas that demonstrated usefulness for many stakeholders, but at the same time require better access to microdata in order to better design specific (micro)economic measures. Table 5.3 lists some such basic types of administrative and micro data in Croatia, according to the sectoral approach primarily, but also according to the approach adopted in extreme situations (COVID19). These case studies can be important for researchers, especially as a suggestion for improvement.

Improving open access to microdata for researchers in Croatia requires logistical, legal and ethical considerations to ensure that researchers can access high quality data while maintaining the privacy and consent of the individuals whose data is being used. Below are some strategies by which to improve open access to microdata in Croatia: (1) anonymisation of data (removing direct identifiers, aggregating data and using pseudonyms), (2) data management plans (developing clear data management plans outlining how data will be collected, processed, stored and shared), (3) legal frameworks and licensing (creative Commons licenses, for example, can provide a standardised way to share data), (4) ethical guidelines (ethical guidelines that guarantee the responsible use of microdata), (5) data repositories and platforms (robust data repositories and platforms that should be both user-friendly and secure and should support metadata standards to facilitate data discovery and use), (6) training and support (how to access, use and interpret microdata responsibly), (7) as mentioned above, collaborative networks (between researchers, institutions and data providers), (8) funding (grants, recognition and rewards for researchers who contribute high-quality data to open access platforms), (9) interoperability standards (interoperability standards to ensure that microdata from different sources can be easily integrated and compared) and/or (10) open data policy (sharing of microdata collected with public funds). By addressing these areas, stakeholders can improve both the accessibility and usability of microdata for researchers, thereby enabling more robust and impactful research to be undertaken, all while protecting the privacy and rights of individuals.

5 Conclusion

Each type of microdata serves different research purposes and can provide valuable insights into various aspects of economic, social and health-related phenomena. Access to diverse and robust microdata is essential for conducting thorough and high-quality research. The collection of microdata in EU countries and Croatia usually follows strict guidelines so as to ensure accuracy, consistency and data protection. Croatia's administrative and microdata systems have evolved significantly, driven by EU integration and alignment with international standards. The transition from a

Table 5.3 Case studies on specific (and useful) approaches to the collection and processing of administrative data in Croatia

Example and institution: OVI index (online vacancy index) is a measure of labour market demand, based on online job advertisements, compiled by the Croatian employment service (CES) in collaboration with the Institute of Economics, Zagreb (EIZ)
Specificity, aim and impact: • Job vacancies, occupational categories, qualifications required and geographic distribution - Monthly reports are published on the CES and EIZ websites - The aggregated index results are openly accessible to the public, detailed microdata might require specific data access agreements
Example and institution: COVID-19 mobility data on the mobility of residents during the COVID-19 pandemic to gauge the effectiveness and compliance with lockdown measures collected through collaborations between government agencies and mobile network operators
Specificity, aim and impact: • Movement patterns, hotspot analysis and compliance metrics - The aggregated index results are openly accessible to the public, while detailed microdata might require specific data access agreements - Aggregated reports were provided by the Croatian Institute of Public Health (HZJZ) and by partner organisations - Aggregated results are available to the public, with restricted access being granted to granular data for research purposes upon request
Example and institution: Survey of health, ageing and retirement in Europe (SHARE)—Croatia as a longitudinal study that collects microdata on health, socioeconomic status and social networks of individuals aged 50 and over conducted by SHARE-ERIC with data collected by national teams, including by Croatian partners
Specificity, aim and impact: • Health indicators, retirement patterns, mobility and economic status - Results are published in international reports and academic publications - Data is made available to researchers upon application through the SHARE-ERIC infrastructure
Example and institution: Croatian longitudinal survey on Well-being (CROLLO) as a longitudinal study focused on understanding various aspects of Well-being among Croatian citizens conducted by academic or research institutions
Specificity, aim and impact: • Subjective Well-being, mental health, social relationships and economic status - Findings published in academic journals and reports - Microdata is made available to researchers through formal requests
Example and institution: Household budget survey (HBS) collects detailed data on consumption patterns, household income, and expenditures conducted by the Croatian Bureau of Statistics (CBS)
Specificity, aim and impact: • Detailed household expenditures by category, income sources and demographic information - Summary results published in annual statistical reports - Microdata available to researchers through specific application processes managed by the CBS
Example and institution: Labour force survey (LFS) as regular survey providing key statistics on employment, unemployment and labour market participation conducted by the Croatian Bureau of Statistics (CBS)

(continued)

Table 5.3 (continued)

Specificity, aim and impact: • Individual employment status, job search activities, work hours and demographic information - Quarterly statistical bulletins and reports - Microdata can be accessed upon request following approval from CBS
Example and institution: Statistical information on higher education collected on students, faculties, programmes and outcomes in higher education institutions and compiled by the Agency for Science and Higher Education (ASHE) and CBS
Specificity, aim and impact: • Enrolment figures, graduation rates, educational achievement and demographics - Annual reports and statistical bulletins - Access to detailed microdata via formal request procedures for research purposes
Example and institution: The regional development index represents the corrected average of the standardised values of the indicators, at the level of all local and regional self-government units, provided publicly by the Ministry of Regional Development and EU funds (MRRFEU)
Specificity, aim and impact: • Although it certainly has its shortcomings, it is the best "attempt" to monitor the complexity of the development dynamics of all territorials (administrative) levels in Croatia - Results are published on the official web site, but the Ministry of Regional Development and EU funds published the first brochure in which they compiled available data from almost all public (official) sources in the Croatia for the regions (NUTS II and NUTS III level) - The publication is publicly available, but in Croatian
Example and institution: Crobarometer, a public opinion survey that measures the attitudes and opinions of Croatian residents on various topics, including socio-economic issues, political preferences and public policy typically conducted by market research companies like Ipsos pulse
Specificity, aim and impact: • Demographic information, opinions on current events, political party preferences and trust in institutions - Results are frequently published in media reports and government documents - Summarised data is often publicly available, while microdata can be accessed by researchers through specific request procedures

Source Compiled by the authors

custom-made approach, during the pre-independence period, to a more open, integrated, and harmonised system has improved transparency, availability, methodological standardisation, and data quality. Key institutions like the Croatian Bureau of Statistics, Croatian National Bank, Croatian Financial Agency and various ministries all play major roles in this progress. This chapter has detailed Croatia's alignment with EU legal frameworks, thereby enhancing greater access to, and greater protection of, microdata.

Microdata encompasses a variety of datasets, including census data, financial reports, tax records and land registries. While matching these databases can sometimes be straightforward, it may also require specialised programming, particularly for complex datasets like land registries. In certain instances, legal restrictions limit the combination of specific data, thereby necessitating the use of secure environments where in which anonymised data can be analysed. Access to these secure

environments is granted only after institutions ensure the anonymisation of personal (sensitive) data.

Continuous policy and practice improvements are crucial to meeting the growing needs of researchers and policy-makers at the EU level. Efforts focus on enhancing metadata standards, thereby improving data anonymisation and expanding data access. The CBS exemplifies this by adhering to international best practices in data collection, processing and bilingual publication. The CES also plays a significant role by collecting extensive labour market data (Labour Force Survey), including statistics on unemployment, job vacancies and the use of active labour market measures. CES's digital transformation has streamlined data collection and dissemination, thereby improving labour market information availability and reliability but also data, for example, from Income Survey, Household Consumption Survey, Innovation Activities of Enterprises or from the Annual Survey on the Use of Information and Communication Technologies in Households and by Individuals.

Despite these advancements, challenges remain in terms of harmonisation, open access and managing the costs and methods associated with data collection and analysis. Institutional data users, such as the Institute of Economics and the Croatian Employers' Association, have highlighted the need for more accessible and harmonised data sources, particularly in machine-readable formats. The Croatian administrative data ecosystem is aligned with the European Statistical System (ESS), but challenges persist in the collection, processing, analysis and availability of microdata. Aggregate data are typically accessible via web sources and databases in Croatian and English, as well as through individual requests. However, most microdata are protected by GDPR and other legal frameworks that require the use of safe rooms for data analysis, where available. In addition, although, microdata from various systems (health, labour market, contributions and taxes, retirement funds, census, etc.) are generally accessible for scientific research and financial accounting indicators, like entrepreneur accounting reports, are available through FINA for a fee, we also wish to highlight several surviving challenges in Croatia:

- At present, there is no single access point for administrative data, but it is planned for the near future;
- Researchers typically must navigate multiple institutions with varying requirements;
- Access to microdata is primarily granted to domestic scientific institutions and universities, leaving independent researchers and smaller organisations largely excluded;
- Obtaining access involves significant administrative hurdles and often necessitates extensive agreements, functionally deterring many potential users;
- While published data is largely available in English, microdata access is primarily in Croatian, thereby presenting a further barrier to international researchers.

In conclusion, although significant progress has been made in improving data access and quality, sustained efforts are needed to address existing challenges and to fully exploit the potential of these systems for policy-relevant analysis. Ensuring the reliability, accuracy and availability of statistics, while protecting individual privacy,

remains a key objective; this objective supports the broader goal of informed and effective policy-making.

References

Connelly, R., Playford, C., J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. Social Science Research, 59, 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015.

Corrado, G. (2019). Institutional quality and access to financial services: Evidence from European transition economies. *Journal of Economic Studies, 47*(6), 1363–1376. https://doi.org/10.1108/JES-03-2019-0131

Crato, N. (2017). *A call to action for better data and better policy evaluation*, a briefing on the importance of administrative data for social knowledge and policy evaluation, at big data times. JRC Science for Policy Report. European Commission, ISBN 978–92–79-65346-9. https://op.europa.eu/en/publication-detail/-/publication/ad3b0218-f280-11e6-8a35-01aa75ed71a1/language-en, 10.2760/738045

Crato, N., & Paruolo, P. (2019). The power of Microdata: An introduction. In N. Crato, & P. Paruolo (Eds.), *Data-driven policy impact evaluation, how access to Microdata is Transforming Policy Design* (pp. 1–14). Springer, ISBN 978–3–319-78461-8 (eBook). https://doi.org/10.1007/978-3-319-78461-8.

Croatian Employment Service—CES. (2024a). https://www.hzz.hr/en/about-ces/

Croatian Employment Service—CES. (2024b). Monthly Statistics Bulletin, Year XXXVII/2024, Croatian Employment Service, Zagreb.

Dale, A. (2005). Research access to microdata: An attempt to provide a context. *IASSIST Quarterly, 27*(4). https://doi.org/10.29173/iq866

Duncan, G., & Pearson, R. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science, 6*(3), 219–232. https://doi.org/10.1214/ss/1177011681

Fuchs, R., Göllner, T., Hartmann, S, & Thomas, T. (2023). Fostering excellent research by the Austrian micro data center (AMDC). *Jahrbücher für Nationalökonomie und Statistik*. https://doi.org/10.1515/jbnst-2023-0043.

HZZ. (2023). Godišnjak 2022. ISSN 1849–4854, Zagreb.

Müller, D., & Umkehrer, M. (2023). International Data Access Network (IDAN) for sensitive microdata in Humanities & Social Sciences. *1st Conference on Research Data Infrastructure.* https://doi.org/10.52825/cordi.v1i.269

Obadić, A. (2016). Tržište rada. In A. Obadić, & J. Tica (Eds.), *Gospodarstvo Hrvatske* (pp. 203–246). Ekonomski fakultet Sveučilišta u Zagrebu. ISBN 978–953–346-040-6.

Obadić, A. (2020). Ekonomska integracija i tržište rada. In R. Jovančević (Eds.), *Ekonomika i ekonomske politike Europske unije* (pp. 285–318). Ekonomski fakultet Sveučilišta u Zagrebu. ISBN 978-953-346-155-7.

Official Gazette. (2018). Labour Market Act, Official Gazette 118/2018.

Regulation 223/2009 of the European Parliament and the Council. (2009, March 11). (update: 8/6/2015) https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32009R0223

Ritchie, F. (2008). Secure access to confidential microdata: Four years of the virtual microdata laboratory. *Economic & Labour Market Review, 2*(5), 29–34. https://link.springer.com/article/10.1057/elmr.2008.73?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot

Nataša Nater Drvenkar is a Full Professor at the Faculty of Economics and Business, Josip Juraj Strossmayer University of Osijek, specializing in macroeconomics, economic policy, and regional economic imbalances. She serves as a special advisor to the Minister of Regional Development and EU Funds, offering expert guidance on regional development policies due to her scientific and professional expertise, rather than political involvement. Her past role as Associate Dean for International Cooperation enhanced collaboration between academia and industry. She excels in structuring and presenting research findings to deepen the understanding of complex economic issues.

Alka Obadić is a Full professor (tenured) in Economics at the Faculty of Economics and Business, University of Zagreb, Department of Macroeconomics and Economic Development. She is also a visiting professor at many EU universities (the Complutense University of Madrid, the University of the Balearic Islands, the University of Lisbon, the University of Malta, the University of Granada, etc.). Her research focuses on macroeconomics, labor market, employment policies, and sectoral clusters. She publishes in Actual Problems of Economics, Amfiteatru Economic, Technological and Economic Development of Economy, European Review of Labour and Research, Scientific Reports, PLoSONE, etc. As a researcher, has participated in many scientific projects financed by EC, Oxford Analytica, USAID, WB, etc.

Josip Tica is a Full Professor at the Faculty of Economics and Business, University of Zagreb. His research focuses on transition economics, macroeconomics, urban economics, and the housing market. He has published in Economic Modelling and Studies in Nonlinear Dynamics and Econometrics and is the author of two macroeconomics textbooks as well as a co-author of several textbooks on the Croatian economy. He serves as a reviewer for 11 international journals, is the editor-in-chief of Ekonomski pregled, and a member of the editorial boards of Croatian Review of Economic, Business and Social Statistics and Contemporary Economics.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 6
Access and Use of Administrative Data: The Case of Cyprus



Nikolaos Theodoropoulos and Andreas Tryphonides

Abstract This chapter describes the status quo regarding access to and the use of administrative data in Cyprus. Current challenges include resolving data ownership issues, addressing authorities' concerns about data anonymity and misuse and raising awareness about the public value of both research and policy evaluation. Government institutions that accumulate administrative data have made substantial progress in providing open access to aggregated administrative data. There is also sufficient IT infrastructure and human capital to adopt international practices concerning the provision of the underlying microdata to researchers. Better coordination among academics, and sufficient political will to overcome current obstacles, would significantly advance ease of access to administrative microdata. We propose the establishment of an office dedicated to this specific purpose.

1 Introduction

This chapter's purpose is twofold. First, it aims to inform the academic and policy-making communities about the current situation in Cyprus in terms of access to administrative data, whether it be individual or firm-level data. Second, it expresses a willingness to initiate a procedure for formal institutionalised access to such data. In Sect. 2, we provide background information about the type of data that is currently collected and the recent changes made regarding open access to government data. In Sect. 3, we describe the current infrastructure and we discuss data ownership, as well as concerns about privacy and anonymity, in Sect. 4. In Sect. 5, we identify key areas of improvement and provide concrete proposals about how we might make further progress on this issue. We conclude in Sect. 6.

N. Theodoropoulos · A. Tryphonides (✉)
University of Cyprus, Nicosia, Cyprus
e-mail: tryfonidis.antreas@ucy.ac.cy

N. Theodoropoulos
e-mail: theodoropoulos.nikos@ucy.ac.cy

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_6

2 Background Information

Even though a significant amount of data is being collected daily in Cyprus, through the interaction of individuals and firms with various public services, such data are only available in aggregate form and not at the individual level. Substantial progress has been made only recently, in terms of aggregate data, with the launch of the National Open Data Portal[1] in February 2024. The launch of this portal was predominantly driven by the relatively recent EU Directive 2019/1024 on "Open data and the re-use of public sector information".

Various aggregated datasets (1,238 as of April 2024) are now openly available to the public, either through traditional forms (spreadsheets) or through APIs (Application Programming Interface) and span a range of public organisations (91 as of April 2024). This includes domains such as Transportation, Population, Society and Employment, Government and Public Sector, Environment, Energy, Education, Culture and Sport, Agriculture, Fisheries, Forestry and Food, Regions, Cities and Zoning, Justice, Legal System and Public Safety, Health, Economy and Finance, Science and Technology. In addition to the Directive, EU Regulation 2023/138 has defined certain datasets as High-Value Datasets (HVD) that are considered to be valuable for society, the environment and the economy.

We are of the view that this is a significant development that opens the way to increased use of such data by the public, by students and (potentially) by researchers in both the private and public sectors. The benefit and added value of making such data publicly available remain to be seen, but we believe that this first step is an important one to take in order to change attitudes and societal norms concerning the public value of research.

Related to the aforementioned datasets, while the underlying microdata is directly made publicly available for certain cases due to their nature (e.g., information about government procurement contracts or hourly measures of air quality), most of the microdata remain private because they contain sensitive information; to that end, only aggregate statistics are provided. In other cases, the data is aggregated in an ad-hoc manner and could potentially be provided, if required, at a more granular—but still aggregate—level without breaching any data privacy concerns. For example, aggregate data are available at the regional level, but can also be provided at the municipality level. Nevertheless, we believe that the summary and meta-data provided by this new portal are very useful as a point of reference for researchers who are unaware of the types of data collected by governmental agencies.[2] The existence of meta-data points to the existence and maintenance of an underlying database of individual units. For example, providing statistics for undergraduate students by age, year and gender suggests that there is a database that is maintained at the individual level that contains at least this information per student. Furthermore, the portal opens up the possibility to users to make data requests, which are handled by the corresponding authorities that own the data directly. This implies that they are examined on a case-by-case

[1] For more information, please visit https://www.data.gov.cy/en.

[2] The information is provided only in Greek as of April 2024.

basis as though the portal did not exist. This is because it is unlikely that authorities will provide any individual data at a more granular level than the one that is already published in the portal. As far as we are aware, being part of a research institution does not make a difference to the way in which the request will be handled.

In general, access to administrative microdata for research purposes is not institutionalised, despite the existence of laws that allow for potential access. The CY law (N.184(I)/2017) regulates the right to access government data and allows individuals who are interested in this information to apply. Correspondingly, the CY law (N. 143(I)/2021) governs open data and seeks to harmonise Cypriot legislation with the aforementioned EU Directive 2019/1024. However, there is still some ambiguity as to whether access to confidential data can be provided for academic research purposes. The decision on data provision is, ultimately, up to the corresponding organisation to decide in consultation with the Office of the Commissioner for Personal Data Protection. This implies that the examination of the request is made on an ad-hoc basis without transparency in the decision-making process. This, de facto, creates heterogeneity in access. Concerns about data privacy and confidentiality are the top reasons cited for refusing to provide such access.

We are aware of only two academic papers, published recently, that have used administrative data (Clifford & Mavrokonstantis, 2021; Mavrokonstantis & Seibold, 2022). Both papers use the same source of data and study issues related to behavioural responses to tax reforms and tax enforcement. This shows the administrative data's potential value. However, access to the data seems to have been obtained through personal requests to the Republic of Cyprus' Tax Department, and not through a formal institutionalised procedure.

3 Database Infrastructure

There has been important progress in recent years regarding the establishment of the IT infrastructure necessary to provide both the access and management of administrative data in different governmental agencies. The creation of the Government Data Warehouse (GDW) is testament to this and its aim is to provide easy access to accurate and integrated data of the Civil Service and to improve service quality and productivity within the Government.[3] In particular, the data warehouse facilitates access to information from various Government Information Systems. It is structured so as to enable dynamic queries and data analytics; this allows for individual information to be merged across different agencies and this can serve the specific needs of the government. For example, such information is required to establish the eligibility of individuals for National Health Insurance. The fact that there exists an IT infrastructure that already supports such an endeavour implies that many of these

[3] We should clarify that the GDW provides services to the government only, not to the public, and this includes researchers.

government organisations already have the necessary capital in place. GDW can be a potentially pivotal agency regarding access to such data by researchers.

The above observations also apply to the Government Statistical Service (CYSTAT), which operates independently of the GDW. The new Official Statistics Law of 2021 (Law No. 25(I)/2021) has been successful in granting the CYSTAT the right to costless access to data produced by all government agencies.[4] The right to access extends beyond the government domain. For instance, CYSTAT has the right to access data produced by private entities, such as supermarkets and banks. Part of the success has been down to the obligatory nature of data collection, given that such data are required for the development, production and dissemination of official statistics per EU regulations.

Access to administrative data has been granted to CYSTAT by many agencies, including the Population Registry, the Migration Department, the Tax Department, Social Security Services, and the Health Insurance System. This also shows the rapid progress regarding the digitalisation, in at least some parts, of the public service. Unfortunately, other departments work in manual mode, but this is likely to change soon. Another important aspect is that of cross-linkage of datasets so that updates by a particular agency are automatically available to all authorised users. This would allow for a more comprehensive and holistic assessment of policies, especially social assistance ones. However, there exist certain difficulties with matching individuals across databases residing in different institutions or departments, given that they do not use the same identification key across different sources.[5] This is particularly true for the immigrant population.

4 Data Ownership, Privacy and Anonymity Concerns

Although granting CYSTAT or GDW the right to access (potentially different) subsets of administrative data is now a reality, it is not sufficient to make progress concerning the granting of access to third-party researchers for scientific purposes.[6] This is especially true for the CYSTAT, as one would think that this is the most natural mediating agent between the Government and researchers for data access. The reason is that ownership of the data does not reside with CYSTAT, or even GDW for that matter, and hence researchers cannot have access to the microdata through them. The way in which data provision is organised up to this point is that ownership resides with the specific organisation that either generates or collects the data, not

[4] According to the Official Statistics Law of 2021 (Law No. 25(I)/2021), one of its main new provisions is "*strengthening the mandate regarding access to and use of administrative records for statistical purposes*". The right to access is, nevertheless, only granted to CYSTAT in order to fulfil its obligation to produce official statistics.

[5] Some government departments use a person's identity card number to identify individuals, whereas others use the national insurance number.

[6] It is important to clarify that CYSTAT provides data to the Open Data Portal, albeit only in aggregate form.

with CYSTAT or with GDW because these are considered consumers (receivers) of government data in this process. Such an endeavour becomes even more difficult when we consider the experience of researchers in gaining access to the micro survey data produced by the CYSTAT. Data owners are overly cautious about providing access to such data, frequently citing privacy and data protection concerns. What is perhaps less commonly known is that one can make a complaint to the Commissioner for Personal Data Protection, who can make the final decision based on the law and on the specific characteristics of the information requested.

Furthermore, neither institution has the evaluation of public policy under their mandate. This is potentially one reason why there is no incentive to facilitate such access. This also implies that neither institution employs researchers to implement counterfactual analysis or public policy evaluation.

More generally, the current status quo for data access, whether its origin is administrative or survey-based, can be largely attributed to concerns about privacy. While the computerisation of the public service has enabled data sharing within the government, at a technical level, authorities frequently cite the risk of de-anonymisation and, hence, deny overall access when it comes to third-party access and anonymisation of the data. We find these concerns to be largely unwarranted, given that other EU countries have already successfully implemented such anonymisation. This can be explained by lack of information, lack of willingness to obtain information on how to successfully implement such anonymisation, or by selective interpretation of the GDPR when it comes to third-party access.

The international practice involves strict contracts between data providers and researchers; there are clear guidelines and regular checks. This should be sufficient to dispel concerns about data de-anonymisation. A three-stage process is usually followed during anonymisation: First, population samples are drawn. Second, identifying variables, such as names, addresses and social security numbers, are deleted. Third, other variables with sensitive information are also deleted. Even then, the data are not provided directly to researchers and extra anonymisation steps take place through further information reduction (i.e., top or bottom coding and through the deletion of digits of hierarchical variables) and secure modes of access. Finally, results obtained from the data analysis go through disclosure control before becoming publicly available and this guarantees that no persons or firms can be identified. Further processes are required for the transmission and accessing of the data by researchers (see Antoni & Schmucker, 2019 and the references therein for the standard procedure followed by the German Federal Employment Agency at the Institute for Employment Research (RDC-IAB)). One could also resort to internal staff implementing the statistical analysis using software and programs pre-coded by the researchers in cases in which information sharing is strictly prohibited.

Overall, authorities claim that they seek a balance between the public good of providing such data and the risk of breaching privacy. In terms of this balance, we find that the benefits are undervalued while the risk is overstated; this leads to a very conservative outcome. We should also stress that local researchers have successfully utilised administrative data provided by non-Cypriot entities and understand both the

procedures and the consequences of breaching the confidentiality terms of contracts with data providers.

Reflecting on the above observations, one can easily conclude that Cyprus has both the necessary infrastructure and human capital to make the leap and to achieve the same level of maturity in data access as other European countries. What is perhaps missing is sufficient trust in government procedures and the necessary political will to implement this change. Change must come from the top, since most of the civil service has a top-down management approach.

5 The Way Ahead

5.1 Coordination

The infrastructure that already exists in Cyprus is sufficient to supporting a change in regime. Escaping the current status quo requires some experimentation concerning ad-hoc projects that could show the potential gains yielded by sharing data. One such case of this is the recent cooperation of the Economics Research Centre of the University of Cyprus (CyERC) with the Tax Department for a project on tax reform. Although access to the actual microdata was not granted to the CyERC, aggregate information that was not otherwise publicly available was shared by the Tax Department. Further collaborations, based on bilateral agreements for sharing data, could be beneficial for fostering a transition toward a more fertile research environment. More generally, the experimentation stage could be framed as a "piloting stage" in which incremental steps could be taken. Established researchers could be trusted with specific datasets and the research outcome could be communicated to the respective government authority. This could also, of course, include an evaluation of a specific public policy.

Despite experimentation being a supportive initial step in this transition to a new status quo, permanent change will follow once the procedure for data access is institutionalised. A coordinated effort by the research community could apply some pressure on the government; however, ultimately, the decision is political. In order for it to become viable, it should be directed from top management and by individuals who understand the public value of such data.

Since administrative data already exists for Cyprus, in some form or another, and not many extra resources are needed to generate such data, we propose the introduction of an office, akin to Research Data Centres (RDCs), like those that already exist in other countries (e.g., Germany (through the Institute for Employment Research), UK (through Administrative Data Research UK) and Canada (through Statistics Canada)). The scope of this office would be to coordinate the different governmental institutions for data linkage and to acquire the necessary technological infrastructure to securely source the data from all of the different institutions. Moreover, office employees should prepare and organise the data to be used for research purposes,

produce meta-data and facilitate access for researchers and research institutions. They should also provide sufficient data documentation, comprehensive guidelines and relevant training to eligible researchers.

We believe that it would be efficient for this office to administratively belong to either CYSTAT or to a new government body, provided that the data ownership issue is resolved.[7] This will go a long way toward solving the issue of data access; the technical infrastructure and know-how also already exist and could be substantially enhanced with further resources and employee training. This office should work in close cooperation with universities and with research institutions and its major role would be to monitor and facilitate data access. More specifically, it should guarantee that the data is used safely and ethically and that any evidence cannot be tracked back to either individuals or organisations. Close collaboration with the Cyprus National Bioethics Committee should ensure the upholding to ethical standards. Moreover, certain eligibility criteria could be common to all kinds of datasets while they should be stricter for other datasets. For example, certain information could be restricted to researchers who are based in Cyprus or who are affiliated with a Cypriot institution.

Another role that the office might play would be to monitor the demand for administrative data through user surveys and to make queries for preferred modes of access and types of data. It could organise user workshops in which researchers and administrators could exchange ideas about best practices, and conferences where researchers could promote their findings.

Funding for such an office could be provided through the structural funds of the Research and Innovation Foundation, other government funds, cross-universities budgets,[8] or through local or international research funds that researchers might hold.

5.2 Changing Attitudes

Despite some agencies' willingness to collaborate in disseminating administrative data, the legal framework for access to administrative data is relatively new; hence, the agencies have several functions and responsibilities to which to adhere. Furthermore, they are also fairly resource-constrained and understaffed. So, while they see the advantages of providing these data, there are substantial implementation lags.

There are also quite reasonable concerns around ethical issues, such as providing to researchers the data that were not originally collected for research purposes, but for doing things like paying taxes or providing welfare benefits. Academic institutions should pursue an organised effort to convince data providers about established procedures by which to guarantee data anonymity. This campaign should also be extended

[7] According to international standards, RDCs administratively belong to the National Statistical Centres/Bureaus. Nevertheless, it is not clear whether this is possible in Cyprus under the current legal framework that defines the responsibilities of CYSTAT.

[8] Despite its small size, Cyprus has three public universities and nine private universities. Both public and private universities could contribute to such a fund.

to the rest of the public sphere as politicians might not necessarily understand the full extent of the social benefit offered.

5.3 Value Added

If our proposition for the establishment of such an institution or an office within an existing institution proves to be fruitful, then the potential for academic research in Cyprus and for evidence policy-making would be substantial. It would also allow researchers to make comparisons between Cyprus and other European countries in terms of policy effectiveness and outcomes, to learn from other countries and to identify potential areas of improvement.

Having administrative data, for instance, would help to formally evaluate the various active and passive labour market policies that Cyprus' government is operating through its Human Resources Development Authority (ANAD). Furthermore, a national minimum wage was recently introduced in Cyprus. Using administrative income tax data, or perhaps even the actual employment forms filled out by employers, would allow researchers to not only understand the incidence of low pay, but also to examine what happens to low pay workers over time and who moves in and out of low pay. Moreover, Cyprus has been home to a large number of immigrants and refugees because of the European enlargement and the ongoing wars and conflicts in Ukraine and the Middle East. Having administrative data on the number and the characteristics of immigrants and refugees is paramount to examining their length of stay and to their assimilation. Another interesting area of topical research, which could be made possible by gaining access to administrative data, is evaluating the cost-effectiveness of the new National Health System (GESY) in Cyprus and the welfare effects to households through its introduction. The above research questions can only be answered by following the same individuals or firms over time.

Up until the present, and to the best of our knowledge, there has not been any research on the impact of the four policies suggested above. Thus, a lack of access to administrative data precludes researchers from providing thorough and causal evidence regarding their effectiveness. Such evidence would greatly help the government to improve policy-making.

6 Conclusion

Cyprus seems to be behind the curve when compared to other European countries. Yet, we have identified several positive aspects of the current status quo. First, the legal framework for open access to data has at last been substantially revised. While this does not solve the issue of access to administrative data by researchers, it remains an important first step. Second, there is a willingness for cooperation from some institutions that hold administrative data. The information technology infrastructure and

human capital already exist. This implies that Cyprus can easily adopt international practices. Moreover, there is significant demand from the academic community for access to administrative data.

At the same time, we have identified several challenges that need to be addressed urgently. First and foremost, the issue of data ownership should be resolved. This step requires bold decisions by the government. Second, there are misplaced concerns about data anonymity and trust issues regarding data misuse. This makes the introduction of a formal framework for data access a necessity. Third, government institutions that have made progress in gaining access to administrative data work individually, rather than collectively. Greater coordination between them should increase efficiency and could potentially self-generate the resources required for the establishment of an office that manages access for third-party researchers. Finally, there is no unified effort from the academic community to inform politicians about the usefulness of administrative data. Efforts for access to administrative data have been done individually to date and with minimal success. Greater coordination on the researchers' side will raise awareness of the public value of administrative data and might, potentially, convince the government to implement the changes required to make progress on this issue.

References

Antoni, M., & Schmucker, A. (2019). The research data centre of the German federal employment agency at the Institute for Employment Research (RDC-IAB)—Linked microdata for labour research. *International Journal of Population Data Science, 4*(2), 05.

Clifford, S., & Mavrokonstantis, P. (2021). Tax enforcement using a Hybrid between self– and third–Party reporting. *Journal of Public Economics, 203*, Article 104519.

Mavrokonstantis, P., & Seibold, A. (2022). Bunching and adjustment costs: Evidence from Cypriot tax reforms. *Journal of Public Economics, 214*, Article 104727.

The Official Statistics Law of 2021 Law No. 25(I)/2021. Retrieved from https://library.cystat.gov.cy/NEW/Official_Statistics_Law(unoftrans)-2021-EN-290421.pdf

Nikolaos Theodoropoulos is Associate Professor of Economics in the Department of Economics at the University of Cyprus and runs the Survey of Health, Ageing and Retirement in Europe (SHARE) for Cyprus. He has worked at University College London (UCL) as Senior Research Officer at the Centre for Research and Analysis of Migration (CReAM). His research has been funded by the Economic and Social Research Council, the Boston College Centre for Retirement Research, the U.S. Social Security Administration and the European Commission. He has published in journals such as the European Economic Review, International Journal of Industrial Organization, Journal of Economic Behavior and Organization, Journal of Economics and Management Strategy and the American Economic Review (P&P).

Andreas Tryphonides is Assistant Professor in the Department of Economics at the University of Cyprus. He holds a Ph.D. (2017) in Economics from the European University Institute in Italy,

where he also earned an MRes in Economics (2012). He holds an MSc in Econometrics and Mathematical Economics (London School of Economics) and a BSc in Economics with Econometrics (University of Kent). His research focuses on applied and quantitative macroeconomics and econometrics, which includes leveraging information contained in micro and aggregate data to address macroeconomic questions and related identification issues. He has published in renowned journals like the Journal of International Economics and the Review of Economic Dynamics.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 7
Learning from Administrative Data in Czechia: Hiccups at the Last Mile

Petr Bouchal and Dominik Hajduk

Abstract Access to administrative data held by Czech public bodies is not currently governed by specific rules. The state's data architecture establishes a set of principles, registers and identifiers that provide a good basis upon which to both manage and link data; however, this is neither used in practice to enable the reuse of administrative data for the purposes of research or evaluations, nor are there data catalogues or a clear administrative procedure for requesting access. Access procedures and safe rooms exist for certain statistical data and in the health domain, albeit to some extent. In practice, permission to use data can occasionally be negotiated on an ad hoc basis (this chapter provides concrete examples of this). Some detailed, non-personal administrative data is published for transparency purposes and has been used for high-quality academic research. A new draft law concerning access to administrative data would alleviate these gaps, but has stalled in the legislative process at the time of writing.

1 Introduction: State of Play

Much of the current state of the accessibility of administrative data in Czechia is a consequence of a transparency and open data drive dating back to the 2010s. This means that the use of data for research is an afterthought in many policy areas and that existing data releases mostly include either aggregate data or fiscal and financial data relevant to journalists and for the purposes of watchdog activities, rather than for policy-relevant research. An exception is "data on the state," which is published in high levels of detail for transparency purposes and which does not entail privacy and confidentiality concerns—notably data on public spending, procurement, and contracting that have been used in cutting-edge research.

P. Bouchal (✉) · D. Hajduk
Government Analytical Unit, Government Office of the Czech Republic, Prague, Czech Republic
e-mail: petr.bouchal@vlada.gov.cz

D. Hajduk
e-mail: dominik.hajduk@vlada.gov.cz

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_7

Similarly, the basic architecture of data systems that determine how data might be linked and accessed is largely shaped by decisions made concerning the entire state's information architecture from either before or around 2010. This brings with it a legacy of large, typically outsourced IT systems containing data assets which are sometimes not well understood, even by the public institutions responsible for them. At the same time, the system of registers and identifiers that lie at the centre of the state's information architecture, where implemented properly, enables data linkage, even if primarily for service delivery, and provides a bedrock for ongoing efforts to enable data linkage and accessibility for research. However, a large gap between aspirations and the reality of both data management and cataloguing hinders effective data discovery, linkage and reuse in service delivery, as well as for the purposes of research. There are also legislative and organisational gaps that make accessing data difficult for the purposes of research, though these are now being addressed through the establishment of the Digital and Information Agency and through a proposed Act on Data Management and Controlled Data Access.

The current institutional setup also rests on a long-lasting division between administrative data and statistical data. The Czech Statistical Office, as the main statistics authority, focuses squarely on producing statistics, whereas bodies that hold administrative data pay only very sporadic attention to researchers' needs. A researcher will notice this in the absence of a central authority or information point through which they might inquire about existing data sources or receive guidance or even technical support. To the extent that there are facilities for the reuse of public sector data for research, these are established by the statistical and sometimes administrative bodies with little involvement from wider academia—unlike in several European countries in which academic consortia have been established to build up ready-for-analysis linked datasets (some examples of which are included in later sections of this chapter).

Overall, this means that:

- Gaining access to administrative microdata requires the researcher to understand where that data is—some catalogues do exist, but are very partial;
- Ad hoc agreements are typically made for data access, with relatively little predictability as to how data protection will be weighed against the benefit of making data available;
- No clear, legal and practical framework or technical infrastructure currently exist by which to link data from multiple sources.

This chapter proceeds as follows: in Sect. 2, we provide a brief description of the governance of statistical and administrative data. We also present three short case studies that highlight existing research that utilises administrative data from Czechia to illustrate the broad dynamics of data access for research. In Sect. 3, we present a thematic overview of various administrative data sources, though the decision of whether to grant access or not can often seem arbitrary (as there are few formalised procedures, as noted above), and the general expectation often means that this data is readily available only for its original administrative purposes. In Sect. 4, we offer additional commentary about horizontal issues that may prove valuable to researchers

interested in Czech administrative data. Section 5 concludes with a broad summary of current trends and highlights existing and prospective initiatives that aim to both improve administrative data access and utilisation.

2 The Data Landscape

2.1 The Statistical System

The statistical system is legally and institutionally separate from the administrative data landscape. It consists of the Czech Statistical Office (CSO) and dedicated units within ministries, regulators and some specialised agencies, such as the Institute of Health Information and Statistics (IHIS). Its data collection and products are governed by an (annually updated) bylaw; namely the Statistical Survey Programme.[1]

Statistical microdata can in principle be accessed for research purposes via the CSO's Safe Centre, which provides basic technical infrastructure and support for both data linkage and cleaning. There is a formal process through which to apply for access on a per-project basis; the CSO also checks the outputs so as to prevent unauthorised disclosure of protected information. There is currently some uncertainty as to who is determined to be performing research, as opposed to conducting adjacent activities.

There is currently still some fragmentation of the access channels for aggregated statistical outputs. Some data is provided as open data,[2] some through a database interface and some is only available as tables that are downloadable from the CSO's website. There is also a publicly accessible information services channel via email that is generally responsive to requests made regarding, for example, custom aggregations.

As the statistical system increasingly relies on administrative data, most notably for running the census and rolling census, it is becoming the de facto path through which to both access and link administrative data for research purposes. The CSO already reuses data from administrative bodies, cleaning and linking them internally, and it may be possible to access these data under specified conditions through the CSO Safe Centre in some cases.

2.2 The Architecture and Governance of Administrative Data

Administrative data's structure, discoverability and reuse potential is largely determined by how it is managed and governed in individual IT systems that are, in turn,

[1] See, for example, the 2025 Programme at https://csu.gov.cz/vykazy/program-statistickych-zjistovani-na-rok-2025.

[2] An R package is available to access this data directly, see Bouchal. (2025).

regulated by a set of requirements that individual institutions comply with to varying degrees.

At the core is a set of so-called core registers (e.g., of persons and legal entities with identifiers) and these cannot be accessed directly from outside, but are instead used to store core information that is then reused by individual digital services that employ a system of identifiers.

Notably, there is also:

- A register of IT systems designated as public administration information systems that provides a public listing[3];
- A register of public services and their associated processes, information systems and data that is being processed for the purposes of providing that service; this metadata can be accessed publicly.[4]

The information from these lists is rudimentary, often incomplete and is aimed at technical experts seeking to facilitate data sharing for public services rather than at researchers, but can be a good starting point for understanding which data is held by which institution.

There is, however, neither a comprehensive catalogue of administrative data nor is metadata publicly available for individual databases held in a standardised or centralised form.

Beyond this, there is a solid system of data dictionaries for published datasets, most of which are managed by the CSO and are provided as open data, for the most common categorisations and units at least, such as geographical and administrative units.[5] Some data providers unfortunately create their own categorisations for the same phenomena, but this is improving.

Open data is provided for some administrative data sets, but this is rarely true for individual records; the most granular data is typically provided for transparency-related purposes on administrative procedures, such as public procurement, contracts, subsidies and public finance more generally. Most open data is catalogued in the National Open Data Catalogue and can be found at data.gov.cz.

Spatial and geographical data is governed somewhat separately, but is generally well managed and provided publicly with good quality and accessibility standards. The primary provider is the Czech Administration of Land Surveying and Cadastre,[6] who also manage the canonical spatial register (RUIAN) that identifies locations and relations down to individual addresses.

[3] https://rpp-ais.egon.gov.cz/AISP/verejne/isvs/zobrazeni-isvs.

[4] https://rpp-ais.egon.gov.cz/gen/agendy-detail/.

[5] https://apl2.czso.cz/iSMS/klaslist.jsp.

[6] https://geoportal.cuzk.gov.cz.

2.3 Recent Examples

Below, we describe three research outputs based on accessing and linking administrative data. They demonstrate what is possible in areas in which data is of a particularly high quality (Structure of Earnings) or where it has long been provided for transparency purposes (subsidies, public procurement).

2.3.1 Estimating the Public Sector Pay Premium

The Information System on Average Earnings (ISPV) is a particularly high-quality data source that is chiefly used to produce earnings statistics. It comprises administrative data provided by all public sector bodies for all employees and of a stratified-sample survey of private sector entities. It provides a detailed breakdown of earnings alongside detailed demographic characteristics and job-specific information, such as ISCO codes.

Bolková et al. (2025) use this data to estimate the public sector pay premium by segmenting the labour force into several hundred segments that are comparable between the private and public sectors. The paper provides valuable information for policy makers concerning the distribution of the pay premium across professions, regions and education levels, pointing out the outsized pay gap that is experienced by highly-qualified professionals in the public sector, particularly in Prague. It also identifies some limitations of the data, from sample size problems on the private sector side to issues with the implementation of the ISCO's classification and the problem of shadow employment via sole-operator companies that "hides" workers from labour market statistics, particularly in some sectors.

2.3.2 Investigating Fairness in Public Procurement

Baránek and Titl (2024) use two transparency-focused data sources—on public procurement and elections—to shed light on whether or not political connections skew public procurement outcomes. Similar work has been undertaken that links political finance donations to procurement outcomes (Titl & Geys, 2019). Procurement and contract data is generally publicly available and a rule stipulates that the unpublished contracts are invalid which ensures completeness.

2.3.3 Evaluating the Effects of R&D Subsidies

The situation is somewhat more complicated with regard to subsidies. Data is generally made publicly available on the recipients of public subsidies; a notable comprehensive source is the register of EU funds recipients.[7] However, rarely is there public

[7] See https://dotaceeu.cz/cs/statistiky-a-analyzy/seznam-operaci-(prijemcu).

information on unsuccessful applicants and this limits what research designs can be used. Consequently, high-quality research in this area depends upon detailed cooperation between researchers and public authorities who award the subsidies; these provide data on unsuccessful bids, typically on a confidential basis, and often also facilitate linking subsidy data to business registers and to economic data about the companies affected. Specifically, Bajgar and Srholec (2025) use data on a selected R&D subsidy programme, run by the Technology Agency linked to a statistical survey of R&D performing entities, to provide evidence on the effects of these subsidies.

3 Sources of Administrative Data and Access to Them

As noted above, the general condition is that data is unlinked, not catalogued and its stewards do not expect the data to generate research interest. Aggregate statistics are occasionally published incidentally, as spreadsheets or PDF files, on the institutions' websites.

In what follows, we aim to help interested readers understand which institution holds data in a particular area and what the current state of play is with respect to the reuse of such data for the purposes of research. We expect that this work will be done in greater depth and will have better information if/when the Act on Data Management is implemented (see final section).

3.1 Population, Security, Migration

Information on **foreign nationals**, such as their names, addresses, nationalities and documentation, but also any records about, for example, their contraventions/infractions under Czech law, are recorded in the Foreigner information system, which is administered by the Ministry of the Interior and by the Czech police.

Aggregated data is also publicly released from the central registry of **firearms**, which is also administered by the Ministry of Interior. Specifically, information about both the holders of licenses (by gender, type, region) and recorded firearms themselves is made available and is updated monthly.

3.2 Businesses

The Ministry of Finance oversees the Administrative Registry of Economic Subjects ("ARES"),[8] which is the source of the key authoritative dataset on legal entities in the Czech Republic. It combines several other registries (e.g., taxpayer, commercial,

[8] https://ares.gov.cz/.

self-employment) and allows for bulk exports as open data, thereby providing a complete overview of identifying information for legal entities.

The Czech Statistical Office administers annual data collection that is then used by other public institutions. One such case is the annual return, provided by economic entities, and which contains **financial information from companies** in accordance with EU regulation 2019/2152 on European business statistics. This data is also shared with the Ministry of Industry and Trade who publish aggregate data about manufacturing industries.

The Statistical Office also plays a key role in **tourism**, where its statistics are a key source of information that is used by public institutions. It administers regular[9] reporting by all mass accommodation providers with regard to both their capacity and occupancy. The results are among the datasets for which the Statistical Office most commonly provides microdata to researchers.

All Czech economic entities are required by law to submit **financial statements**, as per Act No. 563/1991 Coll. on Accounting, which then become part of their records in the commercial registry. From there, these documents are made publicly accessible by the Ministry of Justice.[10] While many companies do not comply with this obligation, or only do so with delays,[11] it can nevertheless be a very accessible and detailed source of data about individual companies.

From 2016 to 2023, the reporting of each sale was mandatory, so the state collected transaction-level data on everyday sales during that time span. It is not clear what the status of this data is with respect to its management and reuse potential.

3.2.1 Research and Innovation (Companies)

The Statistical Office acts as a clearinghouse for data related to **commercial research and innovation**. While it collects some data using its own surveys or reporting mechanisms (based on international best practices, e.g., the OECD Oslo Manual in the case of company innovations), it also compiles other data sources, such as patent data from the Industrial Property Office. While aggregate outputs from these sources are published on the Statistical Office's website, researchers can request access to individual (pseudonymised) records in its safe computing centre.

We comment on public support for R&D in the section on subsidies below. The Ministry of Finance runs the R&D tax relief programme and firm-level data has been used by OECD researchers for a cross-country study.[12]

[9] These are published monthly for hotels and quarterly for other types of mass accommodation providers.

[10] https://or.justice.cz/ias/ui/rejstrik.

[11] While the law contains provisions to punish non-compliance, these are only exercised rarely.

[12] OECD (2020).

3.3 Labour Market and Social Affairs

3.3.1 Labour Market

The Ministry of Labour and Social Affairs publishes aggregated labour market data as open data, including overviews of reported **vacancies, jobseekers and foreigners in the labour market**. Vacancy-level data is published for mandatory vacancy listings at the Labour Office. Additional systems, from which no open data products are generated, include labour inspection records or data on the types of contract relationships.

The Ministry and its subsidiaries oversee further **record-level data**, but there is no legally binding procedure by which researchers would be able to request access. Moreover, the various databases are not linked, meaning that even studies using administrative microdata may not be able to draw complex conclusions. An example of this is a 2024 study by researchers at the Ministry-established Research Institute for Labour and Social Affairs that looked at the causes and evidence for leaving unemployment, but could not link their records to previous or future employers (Hora & Suchanec, 2024).[13]

A high-quality dataset on earnings is the Information System on Average Earnings (ISPV); this includes a census for the public sector and a stratified sample for the private sector, thereby covering 100% of large enterprises. Statistical indicators are released for earnings data, including quantile values of wages by demographic categories and ISCO professions in both salary (public) and wage (private) sectors.[14] An experimental R package is available to access the published aggregates directly.[15] Specific aggregates can be requested at a cost and Trexima, the company that manages both the data collection and processing on behalf of the ministry, has been known to collaborate on research using individual data. Records in this dataset cannot easily be linked at the individual level to other datasets, and ISPV will be discontinued in the coming years as a new integrated labour force reporting system comes into force.[16]

The Ministry of Labour and Social Affairs also provides data on **active labour market policies** and their funding, providing information about the amount and type of support, as well as information on direct recipients (both natural and legal persons). This data[17] stretches back a decade and is updated daily, and as such provides a good example of some of the high-quality administrative data that is published as open data.

[13] Hora and Suchanec (2024).

[14] https://www.ispv.cz/.

[15] https://petrbouchal.xyz/ispv/.

[16] https://www.odok.gov.cz/portal/veklep/material/ALBSDBKMRXKC/.

[17] https://data.mpsv.cz/web/data/aktivni-politika-zamestnanosti.

3.3.2 Living Conditions

The Czech Statistical Office can provide access to anonymised records from the EU-SILC survey to researchers interested in analysing **household living conditions**. This "national" avenue can provide a faster and more responsive way to access microdata for researchers not affiliated with institutions already recognised by Eurostat as research entities. Furthermore, the Statistical Office has been continuously undertaking **household expenditure surveys** since 1957, data from which can also be provided to researchers. Finally, researchers also regularly apply for microdata from the **2021 census**, which includes information concerning on commuting, employment or housing conditions. While these are sources of survey (rather than administrative) data, the mode of access via the CSO will be the same. EU-SILC does not contain identifiers that would allow linking to individual observations in administrative datasets.

Substantial data, including information about income, is collected in the process of administering the **social security system** that currently[18] consists of several payments with different data footprints (i.e., only some are means-tested or include the submission of additional data by recipients). Data on social security insurance mechanisms (sickness insurance, pensions and social security premiums) is collected and used by the Czech Social Security Administration ("ČSSZ"), whereas benefits and allowances (including parental/child allowances) are administered by the Czech Labour Office ("ÚP"). Work is currently underway to both streamline and collect data from these disparate sources at the Ministry of Labour and Social Affairs.

Data on the phenomenon of the **over-indebtedness of households**—a key public policy topic—is the subject of two different data regimes. Information about legally sanctioned debt collection proceedings is compiled in a database that is operated by the professional association of third-party debt collectors. Although nominally operated in the public interest, the database is closed off to both policymakers and researchers. If the situation escalates and households choose to enter insolvency proceedings (receiving a debt discharge after several years, while adhering to a court-mandated repayment plan), then the data about this process is collected by the Ministry of Justice in the ISIR system.[19]

Public institutions are also responsible for ensuring the availability of **social services** and administrative data and is, thus, generated both in the process of accrediting providers and in the performing of services by public sector workers. In the former case, data from the Registry of social service providers is made available as open data on the website of the Ministry of Labour and Social Affairs website, including information on any inspections or details of the services they provide. With

[18] The Ministry of Labour and Social Affairs has proposed merging most current social support schemes into a single, tiered payment. As of May 2025, the proposal had been adopted as legislation and is scheduled to come into effect in October 2025.

[19] https://isir.justice.cz/isir/common/index.do.

regard to social services, the Ministry collects data on engagements via a standardised form, although this is not made available, due to the sensitive nature of the data, and not all engagements are recorded in this way.

3.4 Justice and Criminal Policy

A large cache of data is collected by the Ministry of Justice from **judicial proceedings**, given that both courts and public prosecutors report on their activities on a quarterly basis. Aggregated/statistical data is made available publicly on the infoData page of the Ministry's website.[20] The Ministry is known to allow access to individual records to researchers and supports such secondary use of its collected data, even though there is no standardised procedure by which to do so. The Ministry also administers the **criminal register** that contains information about all persons with criminal convictions and also those with suspended prosecutions.

Spatially indexed data is published about the incidence of reported crimes in a crime map.[21]

Data on **inmates, their sentences and prison stays** is held by the Czech Republic's Prison Service, the mandate of which includes penological research. As such, the organisation collaborates with research organisations (including the Ministry of Justice-established Institute for criminology and social prevention). For example, it has shared individual record-level data about sentence lengths and prison readmissions to researchers in 2024.[22]

3.5 Health and Healthcare

Data from the **healthcare** domain is collected and governed by the Czech Republic's Institute of Health Information and Statistics ("ÚZIS" in Czech), a subsidiary of the Ministry of Health. Its National Health Information System ("NZIS") compiles data from the healthcare service provision registry, from specialised registries (e.g., vaccination, diabetes or mental health), screening systems and statistical reporting.

The ÚZIS is aware of its data's research value and has been taking steps to boost the valorisation of healthcare data to the extent made possible by the applicable legislation (especially Act No. 372/2011 Coll.). Since 2023, it has been working on establishing synthetic data as the primary route for researchers to work with its

[20] The infoData page, available at cslav.justice.cz, follows a very subject matter-specific logic, unfortunately, and is not easily accessible to laypersons.

[21] https://mapakriminality.cz/.

[22] České priority (2024).

data[23]. In Spring 2025, ÚZIS plans to launch a service for registered users to submit SQL scripts for processing synthetic data based on four main sources:

- The register of provided healthcare services;
- Causes of death;
- Oncological registers;
- The pregnancy and neonatal health modules of the reproductive health registry.

ÚZIS can be contacted with ad hoc requests[24] for the analysis of NZIS data. Through this process, ÚZIS staff can share statistical summaries or tables/charts of data across its registries and the process is relatively streamlined and should take no more than 30 days after a sufficiently detailed request has been made.

ÚZIS does not rule out the possibility of granting access to anonymised individual records for scientific/research purposes. However, such access is granted on a case-by-case basis and is subject to strict restrictions meant to minimise the risk of re-identification (e.g., no access to data for small areas or for a small number of healthcare providers).

3.6 Education

The Ministry of Education consolidates records with **personal information about students** enrolled in primary and secondary education throughout the country, although not all information contained in school records is reported to the Ministry. The Ministry's decree No. 364/2005 exhaustively lists the datapoints to be reported. These do not include, for example, the subjects taken, results achieved or pupil conduct assessment—such records are only kept locally.

More substantial, consolidated data is available to the Ministry with regard to **tertiary education**, where it is kept in the "SIMS" system and is updated four times per year. In addition to information about the school and the student, the level, length and subject of the enrolled course is also reported and is, thus, available to the Ministry.

Since 2017, secondary schools offering the standard *Maturita* final examination have taken part in **standardised admissions tests** in the Czech language and in Mathematics. Starting with the 2024 cohort, the testing agency made anonymised individual records accessible with student testing scores and the list of schools and subjects they applied for, as well as the reason for their rejection/admission.[25] The ***Maturita* examination** itself is also partially standardised (in languages—Czech, English, Spanish, French, German and Russian—and Mathematics), and the

[23] See Komenda et al. (2025)

[24] www.uzis.cz/res/file/dokumenty/formulare/zadost-analyza-dat-en.doc.

[25] "Data po uchazečích" (applicant-level data) is available at https://data.cermat.cz/menu/data-a-analyticke-vystupy-jednotna-prijimaci-zkouska/datove-soubory.

results—individual records of the answers submitted and scores for anonymised individuals—are also made available publicly.

Another batch of administrative data that boosts transparency and that is already available publicly concerns **schools and other educational institutions** (e.g., art schools/conservatories, language schools or youth centres). The Czech School Inspectorate publishes a standardised set of information for each school (key staff, as well as use of ICT, language learning, etc.), along with all of the inspection reports.[26]

Data on teachers' and other school employees' salaries is collected as part of ISPV (see above) and the Ministry of Education holds more sector-specific, individual-level records about the remuneration and employment of these staff.

3.7 Elections, Political Finance, Political and Legislative Processes

3.7.1 Elections and Political Finance

All electoral data—from candidates to precinct-level election returns—are managed by the CSO and are published as open data. There are public records of political donations.

Candidates are not uniquely identified across elections, but researcher-led efforts have provided cleaned datasets.[27]

3.7.2 Digitisation of the Legislative Process

The Czech Republic is also currently in the process of revising the digitisation of the **legislative process**. From January 2026 onwards, all new legal acts will have to be drafted, submitted for interdepartmental consultation and reviewed via this new system that will generate linked data about the legislative process.

At present, most key steps of each proposed law or bylaw can be tracked online, including most of its accompanying documents in the public path of the eKlep system.[28] This data is partially accessible via an API. Parliamentary procedures are also published in open formats, for the most part.[29]

[26] https://portal.csicr.cz/Search/School; inspection reports available as unstructured.pdf files.

[27] Linek et al. (2025).

[28] https://odok.gov.cz/portal/veklep/materialy.

[29] https://www.psp.cz/sqw/hp.sqw?k=1300.

3.8 Transportation

The Ministry of Transportation oversees various registries of both persons and objects, most notably the registry of **vehicles** (containing both administrative and technical information about the vehicle) and the registry of **drivers**, which also includes information about the traffic violations recorded for each driver. Some aggregated data may be freely exported and summary statistics are available as open data.

Similarly, vehicle-level data can be accessed on mandatory technical inspections of motor vehicles. Data about traffic accidents is published on a map portal and can be downloaded, albeit only to some extent.[30]

There is generally good-quality spatial data on transportation infrastructure and public transit that can be accessed via the broader spatial data infrastructure that is maintained by the Cadastral Office. Data about public transit is published by most large cities and can be found in the national open data catalogue.

3.9 Environment, Energy and Agriculture

Data about the environment is provided by CENIA, the agency that performs the statistical functions in environment policy. This includes, for example, reporting on the state of the environment, but also on polluters, landfills and contaminated areas.

In agriculture, there is an established set of registers (e.g., a register of animals or farmers, crop land).

Subsidies in all of these policy areas are also reported publicly by each provider and are reflected in centralised transparency data on subsidies (see below).

3.10 Public Finances

3.10.1 Spending, Budgets and Accounts

The Ministry of Finance has noticed a significant increase in the usage of its **public budgets** data tool in recent years.[31] This system allows public users to examine state spending, either from the budgeting or accounting points of view.[32] It provides public access to information from all of the public sector's institutions, both national and local (approximately 18,000 reporting units in total). The available open data exports

[30] https://nehody.policie.gov.cz/.

[31] https://monitor.statnipokladna.cz.

[32] These two dimensions are not linked in Czech law, although some mechanisms such as "reservations" bridge their logics.

are more complete than what is available through the point-and-click interface. An R package is also available for accessing and processing this data.[33]

3.10.2 Tax Collection

The Financial Administration is tasked with the administration of most of the **taxes** that exist in Czech law. Aggregate information about the number of payers and the amounts for various types of taxes in monthly cadence collected, as well as specific information on individual lines of tax returns (e.g., tax losses claimed or various tax rebates in income taxes) is made available publicly. On the individual (person/company) level, however, Czech law has strict provisions about taxpayer confidentiality, precluding research designs that require microdata. Ad hoc agreements between the tax administration and research entitles do exist.

3.10.3 Subsidies and Contracts

All **government grants, repayable assistance and other similar transfers** (including the use of EU funds) are recorded in a database under the Central Registry of Subsidies ("IS ReD")[34] that is operated by the Financial Administration. This aggregates data from various sectoral systems. Exports from this database, allowing for specific analyses (e.g., tagging with respect to individual Partnership Agreement sections or using budget-line categorisations) are made publicly available as open data.

Transparency's key role, as a driving factor behind many of the initial steps in data provision to the public, is further illustrated by the public **registry of government contracts**. Since July 2016, most private-law contracts entered into by public institutions with a value of over 50,000 CZK (approx. EUR 2,000) must be published online, with at least the key metadata machine being readable. This is the source of data that was used in previously noted research, such as Baránek and Titl (2024).[35]

3.10.4 Financial Policy and Regulation

The Ministry of Finance is responsible for administering **support for state-subsidised financial products**, currently including supplementary pension insurance and (building) savings accounts. Accordingly, some information about the financial and living conditions of households is held by the Ministry or can be derived therefrom.

---

[33] Bouchal, P. (2024).

[34] https://red.fs.gov.cz/registr-dotaci/.

[35] Baránek and Titl (2024).

The Ministry also oversees policy on **gambling and lotteries**. It operates a database that contains information about both operators and consumers, including the registry of banned players or data about individual "games" that is updated daily.

4 Additional Considerations

4.1 Data Meaning and Interpretation

4.1.1 Difficult/Non-existent Linkage

Research questions may not be practically answerable with administrative data for instances in which researchers aim to tackle multi-domain questions, such as how education affects labour market outcomes, even if the requisite data exists and has been shared with researchers previously. Crucially, there is very **little experience (and willingness) to engage in linking various datasets**. For example, the Ministry of Labour and Social Affairs only recently launched a project to trial and to establish a programme of the periodic linking of social and health data for administrative purposes.

4.1.2 Metadata and Documentation

Researchers may also come to find out that published or bilaterally shared data is **not properly documented**. This is a consequence of the data being published to comply with transparency requirements (as opposed to research needs) that has driven most data reuse initiatives to date. Specifically, this means that public institutions often understand their primary task to be making data available, as opposed to helping others to understand it. For example, data may be provided with no provided reference as to how to interpret the relation between several files and their structure, as in the case of both public prosecution and court administrative data. Additionally, code lists may often be unavailable or barebones, thereby requiring significant subject matter expertise in order to properly interpret the data.

4.1.3 Statistical Categories

If researchers aim to **compare Czech data** to that of other countries, then they should be mindful of the fact that regional aggregation is usually available at the NUTS 3 level (the "kraj" administrative unit, which is also a central pillar of sub-national government). ISCO and NACE classifications are commonly used, where applicable, in labour market statistics and in other economic datasets, thereby making interoperability easier.

4.2 Accessing and Using Data in Practice

In line with the descriptions above, administrative data holders in the Czech Republic have yet to establish infrastructure for the sharing of (non-open) administrative data to researchers generally. As a consequence, interested researchers should be mindful of potential significant delays, even where both administrative data and the will to share it exist, given that procedures for properly granting access are often developed ad hoc. Two institutions—the Statistical Office and the Institute of Health Information and Statistics—are notable exceptions here, with the former having significant experience in operating a safe computing centre through which to provide access to confidential information to researchers via an air gapped system.[36] The latter is currently establishing similar capabilities of its own, along with other initiatives to promote data reuse, such as the development of synthetic datasets.

It is also good to keep in mind that it is not rare for data to be made publicly available, but can be difficult to find due to not being treated as something that may be of use to researchers. For example, records about the uptake of a scheme to provide extra financing to schools with large shares of Ukrainian pupils following the 2022 Russian invasion of Ukraine were only provided as spreadsheet attachments to announcements throughout the course of the scheme. Therefore, researchers should consider whether there is a "transparency" angle, under which the data that they may be interested in might have been released, and should check for this.

4.2.1 Eligibility, Non-Czech Researchers and Language

Apart from a few exceptions, such as the two institutions mentioned at the start of this sub-section, it is also rare for public institutions to be approached by non-Czech researchers and the availability to either consult or provide support while working with the data might be limited. However, there are no general restrictions or norms in relation to providing data to researchers from other countries.

The Institute of Health Information and Statistics provides English-language application forms; the CSO provides English-language listing of available data and documentation for the Safe Centre, though the application form itself is only available in Czech. Both these official channels (for general statistics and health statistics) require that the application be submitted by a research-performing institution. For CSO, the legal requirement is that the requesting organisation have research as its core mission (interpreted as not necessarily the prevailing activity). For accessing data collected based on European legislation, the CSO may refer to the Eurostat list of research organisations and if needed may generally refer to the national list or research organisations (see below).

There is no generally applicable set of criteria by which to determine whether an organisation is considered as performing research and is, therefore, eligible to

[36] https://csu.gov.cz/operating-rules-of-the-czech-statistical-office-safecentre.

apply for data access. A list of recognised research organisations does exist[37] and it is maintained by the Ministry of Education and the criteria for being listed are somewhat stringent. This is used to determine eligibility for public research funding, for example; however, it is not used explicitly to determine eligibility for data access, but some organisations may refer to it as a decision-making aid.

4.2.2 Cost to Researchers

The CSO lists costs for standardised data products; in other cases, the pricing policies are not generally known, though public bodies may charge fees to recover reasonable costs for freedom of information requests. Both CSO and IHIS will change fees to recover the costs of labour required to both prepare and to provide the data.

Where data is managed by a private contractor—as in the case of the ISPV labour market data—the researcher may be asked to cover the costs of data preparation at commercial hourly rates, based on a contract.

4.2.3 Technical Aspects

The technological means of data access also works on an ad hoc basis because there is no central point for accessing administrative data. The CSO Safe centre is on-premises, but remote access is currently in the works. In other cases, organisations providing data on an ad hoc basis may provide on-site secure infrastructure or might enable remote access, but their approach may be determined by local preferences and by technical circumstances; there is no approach or guidance that is generally applied. This will also determine which software suites are available by which to working with the data.

5 Summary and the Way Forward in Czechia

The current state of administrative data use reflects a system in which data infrastructure is managed in order to provide public services, with some attention being paid to interoperability and to enabling digital provision, but relatively little regard has been paid to the reuse of data as a resource for research and analysis. Neither datasets nor the metadata maintained are catalogued and processes and infrastructure for accessing and linking data is largely absent. As a result, research based on administrative data often results from ad hoc agreements between public bodies and researchers.

This notwithstanding, the core elements of the Czech state's data architecture provide a good basis for the linking of data and for making it available for research.

[37] https://msmt.gov.cz/vyzkum-A-vyvoj-2/seznam-vyzkumnych-organizaci.

Ongoing initiatives, such as the establishment of the Digital and Information Agency with a mandate to oversee data management and data quality, provide hope that data quality and discoverability will improve. Safe room infrastructure for statistical data allows some access to administrative data and research-focused data infrastructure is being established in the areas of health and social affairs.

A legislative proposal that is, perhaps, most important for future prospects in this field and that would help improve both the management of data and the ability to both access and link it for research purposes is currently on the table.[38] The proposed Act on Data Management and Controlled Access to Data would create both specific obligations, with respect to data management by public bodies, and a formalised process for requesting and using administrative data the reuse of which is currently prevented by the prevailing interpretation of national data protection legislation. It would also create a way for public bodies to both link and reuse data held by different institutions.

With respect to data management in specific, the data management would compel a broad range of public bodies who maintain designated information systems to provide a public catalogue and metadata on the data held in these systems, based on a centrally provided set of standards. With respect to data access, the law proposes a formal administrative process through which individuals would submit a request to access individual or linked datasets based on a project proposal to a central point that would then be adjudicated on its merits and feasibility by DIA, as the central authority with reference to opinions of the bodies holding the data (with a right to appeal). Individuals could apply on grounds of research or broader public interest, with the latter providing weaker grounds for accessing sensitive data; institutional affiliation can be taken into account as a guarantee of the researcher's ability to safeguard information, but access is granted in principle to individuals for specific projects. Public bodies would be allowed, by law, to make an inter-agency agreement to link data from each other's information systems for a specific research or analysis purpose. On the technical level, a central safe centre infrastructure could be used to link and provide access to the data, or infrastructure provided by one of the affected data holders could be used. These requirements generally apply to most central public institutions, except for defence and security bodies. Existing access arrangement for statistical data and health data would remain in place and the newly established processes would not apply thereto.

The proposed act has received cabinet approval and has been submitted to Parliament, but is in the very early stages of the parliamentary process. At time of writing, there seems to be a very low probability of the law being passed within the coming months. While the proposed system may be imperfect—e.g., it maintains the division between the world of statistics and administrative data; does not provide for a data research infrastructure in the way other countries do; and provides an administrative process instead of explicit merit-based prioritisation of requests for data access—it still would be a significant step forward because it would make data discoverable

[38] Proposal available at https://www.odok.gov.cz/portal/veklep/material/KORND4KLAAG6/.

and would provide the legal and practical basis for linking and accessing data for research as well as in-house analysis by government bodies.

References

Bajgar, M., & Srholec, M. (2025). Crowding in or crowding out? Evidence from discontinuity in the assignment of business R&D subsidies. *Journal of Public Economics, 245*, Article 105357. https://doi.org/10.1016/j.jpubeco.2025.105357

Baránek, B., & Titl, V. (2024). The cost of favoritism in public procurement. *The Journal of Law and Economics, 67*(2), 445–477. https://doi.org/10.1086/727793

Bolková, P., Doleželová, P., Jurajda, S., Mrázek, P., & Münich, D. (2025). Comparing salaries and wages in the Czech Republic: A bigger problem than it seems. IDEA CERGE-EI, https://idea.cerge-ei.cz/studies/srovnani-platu-a-mezd-v-ceske-republice-vetsi-problem-nez-se-zda

Bouchal, P. (2024). statnipokladna: Use data from the Czech Public Finance Database [Software]. https://doi.org/10.32614/CRAN.package.statnipokladna

Bouchal, P. (2025). czso: Use open data from the Czech statistical office in R [Software]. https://doi.org/10.32614/CRAN.package.czso

České Priority. (2024). Penologická recidiva, gender a další kriminologická data pro ČR. https://www.vscr.cz/media/organizacni-jednotky/generalni-reditelstvi/ostatni/penologicka-recidiva-gender-a-dalsi-kriminologicka-data-pro-cr.pdf

Hora, O., & Suchanec, M. (2024). Postavení starších osob v nezaměstnanosti. https://katalog.vupsv.cz/fulltext/vz_573.pdf

Komenda, M., Tichopád, A., Klimeš, D., Dušek, L., & Válek, V. (2025). Facing the European Health Data Space Challenge: A Synthetic Data-Based Access Model Launched in 2025 in the Czech Republic. Available at SSRN: https://ssrn.com/abstract=5361655 or http://dx.doi.org/10.2139/ssrn.5361655

Linek, L., Škvrňák, M., Šoltés, M., & Titl, V. (2025). Czech political candidate and donation datasets. *Scientific Data, 12*(1), 302. https://doi.org/10.1038/s41597-025-04617-5

OECD. (2020). *The effects of R&D tax incentives and their role in the innovation policy mix: Findings from the OECD microBeRD project, 2016–19* (OECD Science, Technology and Industry Policy Papers. No. 92). https://doi.org/10.1787/65234003-en

Titl, V., & Geys, B. (2019). Political donations and the allocation of public procurement contracts. *European Economic Review, 111*, 443–458. https://doi.org/10.1016/j.euroecorev.2018.11.004

Petr Bouchal is Head of Data Science and Modelling in the Government Analytical Unit of the Government Office of the Czech Republic. In this role, he has led projects using administrative microdata on public sector employment and conducted an impact assessment for legislation on administrative data reuse. Previously, he led analysis and data teams in central and local government and in the private and nonprofit sector. He is the author of several software tools for analytical reuse of Czech public sector data.

Dominik Hajduk is a policy officer at the Government Analytical Unit of the Government Office of the Czech Republic, the central state administrative body. With a background in political science and international relations, he previously worked in a think tank promoting good governance and evidence-based policymaking in the Czech Republic

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 8
Data-Driven Research in Denmark



Maria Schlier and Morten Bennedsen

Abstract This chapter provides an overview of Denmark's comprehensive data environment, highlighting the country's advanced framework for statistical data that is so crucial for socio-economic research. Significant international interest has grown for the structure and content of Danish databases, due to its richness and accuracy and despite the stringent access conditions thereto, due to data sensitivity. Denmark Statistics (DST), the national authority, is pivotal in this context. The chapter details the processes through which to access Danish data, outlines the variety of datasets available and explores the associated benefits and challenges. While primarily focusing on Danish registered data, the discussion also extends to broader data sources, reflecting on Denmark's robust data management practices more generally.

1 Introduction

Denmark and other Scandinavian countries have established robust frameworks for the collection, management and application of statistical data; these data serve as critical resources for a wide range of socio-economic research. At the forefront of these efforts is Denmark Statistics (DST), the national authority that is responsible for data collection and analysis. This chapter provides an overview of Denmark's comprehensive data environment, detailing the processes by which to both access and utilise Danish data, outlining the range of dataset topics available and examining both the benefits and challenges associated with these data resources. The discussion will also encompass broader data sources, even though this chapter's focus will primarily be on Danish registered data.

This chapter is structured as follows: Sect. 1 provides an overview of the data landscape in Denmark, emphasising the role played by Statistics Denmark (DST).

M. Schlier (✉) · M. Bennedsen
University of Copenhagen, Copenhagen, Denmark
e-mail: maria.schlier@econ.ku.dk

M. Bennedsen
e-mail: mobe@econ.ku.dk

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_8

Section 2 details the process through which to access these datasets, including the requirements for institutional affiliation and project approval. Section 3 presents a curated overview of the main types of datasets available and these are categorised per research topic. Section 4 delves into one of the most notable aspects of Danish data: the employer-employee registers. Finally, Sect. 5 concludes and discusses the strengths and challenges of working with Danish data.

2 The Danish Data Landscape

DST oversees the collection of data types that meet diverse research needs; this involves integrating microdata from national registers, survey data and external data sources within its framework. It is an independent government authority within the jurisdiction of the Ministry for Economic Affairs and the Interior. Its structure and functions are based on the Act on Statistics Denmark (Folketinget, 2018). The act puts the institution in charge of the development, collection, production and dissemination of statistical information. At present, 1,800 researchers work with Danish data every month (DST, 2023a).

The backbone of Denmark's data infrastructure is the Central Person Register (CPR), which has assigned a unique identification number to each resident since 1968. This system facilitates the linkage of datasets across various domains including health, education and employment, thereby enabling comprehensive research studies that are characterised by both high levels of reliability and validity. Similarly, the Central Business Register (CVR) assigns unique identifiers to firms, supporting analogous integrations within the business sector. These systems enable researchers to conduct detailed, longitudinal and cross-sectional studies across a variety of topics.

The core of the offerings includes a vast repository of both historical data and recently registered data, all housed in the Research Service's basic databank. The repository covers data about individuals and firms and is updated either annually or bi-annually, depending on the data. The DST conducts surveys that address different population segments and business sectors, including the Danish version of the ILO's Labor Force Survey, thereby complementing the register data.

Additionally, DST incorporates data from other governmental agencies. For example, if a person wants to get access to one of the surveys conducted by the Danish Health Data Agency, then the data request might be processed and requested via the Agency, but the final data transfer and storage are managed on DST servers and, therefore, must adhere to DST's compliance protocols.

Down through the years, while the registered datasets are primarily in Danish, Denmark Statistics (DST) has made efforts to also include documentation in English. Since 2022, the statistical programme has been published in both Danish and English, thereby reflecting DST's commitment to accessibility for a broader audience. These statistical programmes, available in English (DST, 2022, 2023b, 2024) and in Danish (continuously since 2003), provide comprehensive overviews of the official statistics produced by the institution. Additionally, the documentation for various registers

and datasets, which can be accessed at DST's official documentation site, is often bilingual.

3 Accessing Data in Denmark

Accessing the data collections managed by DST is a (long) journey that begins with a formal connection to an authorised Danish institution. Researchers keen on exploring Danish data must first ensure that their affiliation is with institutions that have established agreements with DST. Authorisations to access datasets from Denmark Statistics (DST) are primarily granted to public research and analysis environments, including universities, sector research institutes, ministries and other similar entities. Additionally, research organisations affiliated with non-profit foundations are also eligible to receive authorisations. Authorisation to access datasets from Denmark Statistics (DST) is exclusively granted to Danish research environments, due to the challenges associated with enforcing contracts internationally. However, foreign researchers can obtain access through an affiliation with a Danish authorised institution.

This alignment anchors the host institute (and not only the individual) to the standards of data security. If there is a violation of one of the security agreements, then the punishment can potentially be implemented to the entire institution, not just to the individual who failed to comply.

Once affiliated, the next step that researchers must take is to register themselves through Denmark's Data Portal (DDV). This portal acts as the gateway to the statistical data. The registration process is meticulous and long and involves a thorough review by DST to confirm that the researcher's credentials and intended data usage comply with stringent security and privacy regulations.

Following successful registration, researchers must then write a project proposal. The proposal outlines the purpose, methodologies and specific datasets required for the research. The proposal serves as a blueprint that DST then reviews in order to assess the project's relevance and its adherence to legal and ethical standards. Approval for the project proposal is a formality that must be updated every time that new data enters the server. Researchers must also sign documents explaining that they understand the guidelines. Moreover, all researchers need to pass an annual test to show that they understand the rules on encryption and anonymisation. The detailed procedure to access microdata can be found in DST (2016).

The practical aspect of accessing data also involves understanding the associated costs. These costs vary, ranging from fixed rate agreements for specific datasets to broader framework agreements that span multiple years. The cost for a task involving the Research Service is based on the time it takes to complete the task. As an example, an institution accessing data for the first time that wants to get data on the population, education, income and family registers will be charged between 2,300 euros and 3,150 euros (in April 2024 prices) depending on the institution requesting the data.

Additionally, there might also be costs associated with maintaining the servers over time that are charged on the basis of the number of hours worked by DST.

As mentioned previously, the registered data might be enriched by other data from governmental agencies that is not data from DST. These procedures take longer time and may require (paid) time for anonymisation and processing and, therefore, are more expensive. As an example, setting a server with the Medicines Database and data on demographic conditions from different registers, which is enriched by the Danish Health Data Agency with data from the National Register of Pathology, would cost between 5,000 euros and 7,000 plus the anonymisation and maintenance costs.

Researchers can also conduct nationwide surveys alongside the registered data. These surveys can be administered through various methods including the official digital mailbox, postal questionnaires, telephone interviews or some combination thereof. The results of these surveys can be provided with anonymised identifiers, thereby allowing them to be integrated with registered data. The cost of conducting such surveys varies significantly and is largely dependent on both the survey's scope and the methodologies employed.

4 Data Overview

Denmark Statistics (DST) manages an extensive array of over 440 registered datasets that are detailed in a list known as "*Registre i Forskningsservices grunddatabank*" or in the Register in the research services. Each dataset is accompanied by documentation through the TIMES system, which outlines crucial administrative information, content descriptions, time frames, comparability and availability of the data. However, it is important to note that the TIMES documentation system is available only in Danish at present.

Microdata managed by DST is strictly confined within a secure server environment. It can be accessed remotely[1] from anywhere in Denmark, or from anywhere in the world with an appropriate and secure VPN. Once inside the server, the microdata can be accessed and processed. This data never leaves the premises of DST, except where it has been processed to such a degree that guarantees no individual or firm can be identified under any circumstances. The server lacks internet access, thereby ensuring compliance with privacy regulations. DST provides Statbank for broader public access and this is a comprehensive online repository that hosts thousands of figures and tables containing aggregated data. This resource is readily accessible in both English and Danish. Statbank is also accessible from the microdata server directly, thereby providing a seamless integration of detailed and aggregated data for various research and analysis needs.

Box 1 illustrates a curated, albeit not exhaustive, list of Danish registers and is organised by topic in order to provide a clear overview of the datasets managed by Denmark Statistics (DST). This structured approach helps new users to quickly

[1] Remote.dst.dk.

grasp the range of topics covered by Danish data and to understand what specific information is available. We have also included the specific names of the datasets for more experienced users, given that the repository on DST's website categorises information not by topic, but rather by individual dataset names. While the website provides detailed dataset information, it can be challenging to connect the themes to the appropriate acronyms.

Box 8.1: Danish registers, by topic

Danish registered data
Absence from work: To account for the amount of work that is lost due to absence. It is covered by datasets FRAN and FRPE
Business Enterprise Research and Development: Estimated resources spent in R&D. It is covered by the datasets FUI, FUIK and FUIS. There is also research and development concerning the public sector using FUOF
Car Register and Publications: A diverse set of datasets in use, owners and users of vehicles covered by the datasets DMRB, MOEB, MOSD, MOTP, MOTR and MOTV. There is also data on Road Traffic Accidents covered in FUEL, FUPE and FUUH
Central Population Register: The CPR is the cornerstone of all Danish data. It tracks the main demographic events of every resident in Denmark, including births, deaths, marriages and migration. It is covered by the datasets: BEF, BEFADR, BEFBOP, CIV, CPST, DOD, FAFA, FAHU, FAIN, FLDK, FLYT, FOB, FODT, IEPE, SBDK, STAT and VRDK
Children and young persons placed outside their own homes: Statistics on the municipal authorities' placements of children in out-of-home care. It is covered by the datasets: BUA, BUAF, BUAH and BUAS
Church statistics: Covering people's relations to the national church. It is covered by the datasets: KIRK and SOGN
Crime statistics: There are the datasets on convictions (KRAF, KRKO), Reported criminal offences (KRAN, KRMS, KRSI), Imprisonments (KRIN, KRMS) as well as victims of Reported Criminal Offences (KROF)
Cultural data: Including data on the use of Libraries (BIB) and the Cultural Habits Survey (KVU)
Drug Abuse Treatment: Data related to drug addiction treatments. It is covered in the datasets: SMDB_IBIB, SMDB_VBGF and SMDB_VBGH
General Enterprise Statistics: Covering detailed information on Danish business at the enterprise level. It is covered by the datasets: FIGF, FIGT, FIRE and FIRM
Education: Including data on highest Education Attained (UDDA, UDDF, UDKV and UDSF) Upper-Secondary Education (UDGK), the Student Register (KOET, KOTO, KOTRE and UDG), Primary and lower secondary education (SKOL, UDFK, UDKLASSE_ID and UDSP), Research Education (PHD) and adult education and continuing training (VEUV)
Health Insurance Statistics: Data on the consumption of health care services. It is covered by the datasets: SSHO, SSKO, SSSI, SSSY, SYGL, SYSI and SYST
Hospitalisation: Data on admissions both into and out of treatments in both public and private hospitals. It is covered in multiple datasets, notably SYAM, SYAP, SYHB and SYIN
ICT Use in Enterprises: Covered by the datasets: VITA, as well as the ICT-Expenditure in Enterprises VITU

(continued)

(continued)

Income statistics: Data on the distribution of incomes. It is covered by the datasets: AKM, CONESR, COR, FAIK, INAC, IND, INDAKK, INDFORD, INDSLUT, INHP, INPI and INFU
Interest and Share Price Indices: Contains detailed information about the financial markets. It is covered by the datasets: IRTEPERS, IRTEVIRK, URTEPERS and URTEVIRK
International Trade in Goods and services: Covered in the datasets: UH, UHDI, UHDM and UHDT
Labor market accounts and monthly statistics of employees: On the population's labour market status. It is covered by the datasets AMRTN, AMRUN, BFL and BFLBR
Personal assets and Liabilities: Covering assets and liabilities for different strata of the population. It is covered by the datasets: EJDNR, EJER, EJID, EJVK, FORMAND, FORMBIL, FORMEJER, FORMGELD, PENSFORM, PENSINDB, PENSUDBE, PENSUDBL, PERBOLIG, TOKR, TOS, TPKR and TPS
Persons below the state pension age on public benefits: It is covered by the datasets: BC, OF, OFFL and OFKV
Services to adults with a disability and marginalised adults: Covered in the datasets: HANDIC, HANDICM and Benefit to disability car in HANDICB
Social benefits for senior citizens: Data to display the quality level of municipal services in the elderly care. It is covered by the datasets: AEFB, AEFV, AEHJSP, AELH, AEPB, AEPI, AEREHAB, AERH, AETR and HJSP
Structure of earnings: Covered in the dataset LON and the standardised index of average earnings LONIX_STD
Support for children: There is data about Child Subsidy and Family Allowance covered in BOBO, BOFAMYD, BOTI and BOYD; in childcare before school start covered by BOERNFB; recipients of subsidy for private day-care covered by BOERNPRI; childcare after school start covered by BOERNS and notifications of concern for children and young people in BUU
Unemployment insurance: It is covered by the datasets: AKAS, AMFO and LC
Urban areas: With information on urban areas and the number of people in each urban area and in the rural areas within each municipality. It is covered by the datasets: BBR_ADG, BBR_EJD, BBR_ENH, BBRB, BBRBYGNING, BBRE, BBREJENDOM, BBRENHED, BBRJORD, BYSTRA, BYSTRB and BYSTRS
Asylum Applications and Residence Permits: Covered by the datasets: OPHG, OPHGIN, OPHGPERM and OPHGST
There is also a miscellaneous set of datasets on different topics*: A-Income Statistics:* (INDA), *Banks and mortgage-credit institutions* (REAL), *Board members and managers* (PERSBEST), *Business Demography* (EDRD, EDRN), *Cash Benefits* (KTHJ), *Census of housing* (BOL), *Children and young persons with preventive measures* (BUFO), *Energy consumption in manufacturing industries* (ENERGI), *Environmental Goods and Services* (GROEN), *Foreign Owned Enterprises* (FATI), *Home to work commuting* (PEND), *Manufacturers' Purchases of Goods and Services* (VARK, VARS), *National Elections in Denmark* (VALG), *Purchases and sales by enterprises* (FIKS), *Quality of life indicators* (LIVA), Real property taxes (EJSK), *Sales of real property* (EJSA), *SME Access to finance* (ATF), *Survey on Living Conditions* (SILCD, SILCH, SILCP, SILCR) and *Women's shelters* (KKCOPH, KKCOPHB)

5 Linked Employer-Employee Data

The Linked Employer-Employee Data represents a crucial dataset that is managed by DST. The Integrated Database for Labour Market Research (IDA database) emerged from a substantial effort involving nine to ten man-years of work and was developed to tackle the complexities of tracing enterprise identities over time. This effort was funded collaboratively by the Danish Social Science Research Council and by Statistics Denmark (DST, 2016). Spanning from 1980 onwards, the IDA database provides detailed information about the entire population of both people and enterprises in Denmark. There are multiple high-impact journal articles that have been published by using linked employer-employee data (see Timmermans, 2010).

Initially documented in 1991, the description of the IDA's creation remains available only in Danish (Emerek et al., 1991). However, updated documentation was made available thereafter and this reflects the database's ongoing evolution. The project team, comprised of researchers from Aalborg University, Copenhagen Business School and Statistics Denmark all worked to develop this robust database by using government registers. Since its inception, the IDA database has been updated annually, providing researchers across various disciplines with access to data that connects individuals to establishments and firms, that offers longitudinal insight and that covers the universal scope of Denmark. Bobbio and Bunzel (2018) and Timmermans (2010) have presented a detailed analysis of this dataset.

One of the IDA's notable strengths, in addition to its comprehensive and longitudinal nature, is its capability to be merged with a wide array of existing databases. These can include databases housed at Statistics Denmark, surveys conducted in collaboration with Statistics Denmark or databases from other organisations. Mergers are contingent on the presence of a social security or firm registration number, as these identifiers are crucial for integration. This merging flexibility greatly enhances the utility of IDA for complex research endeavours.

Despite its numerous advantages, the IDA database is not without limitations. Issues arise particularly when analysing data over extended periods, due to occasional breaks in the variables available within the database and the existence of a major break in 2008. These disruptions can make comprehensive long-term analysis challenging. Additionally, the database only recorded detailed information yearly for individuals until 2008, which omits insights into activities that occur between the annual registration periods. To bypass such limitations, there is data available every month from 2008 in another register, the *Labour Market Accounts* which is incorporated into the IDA database indirectly.

These characteristics of the IDA database illustrate both its immense potential for advancing research and the inherent challenges that researchers must navigate in order to leverage its full capabilities effectively.

6 Conclusion: The Strengths and Challenges of Using Danish Data

Denmark Statistics (DST) harnesses data from administrative registers of governmental agencies, thereby ensuring high-quality data coverage of the Danish population, buildings and companies over the course of many decades. This comprehensive data allows researchers to combine data in endless ways and to perform unique analyses of dynamic processes and fluctuations for the entire population of the country.

Each dataset provided by DST comes with an affiliated quality declaration that includes a detailed description of both the data sources and methodologies employed. These declarations detail the quality and content of the data, specifying the purpose, history, collection method, accuracy, reference period and comparability over time.

DST data can be seamlessly integrated with data from other sources, such as survey data or information from other governmental agencies like the Danish Health and Medicines Authority. Additionally, DST can conduct customised interview surveys that are tailored to the specific needs of users, thereby further enriching the data's applicability.

A pivotal dataset within DST's offerings is the linked employer-employee data. This dataset was developed in order to address the complex issue of tracking enterprise identities over time—a challenge that was previously unmanageable for individual researchers due to both the significant time and funding required. The creation of the Integrated Database for Labour Market Research (IDA database), which contains comprehensive information about the total population of people and enterprises in Denmark from 1980 onwards, has been instrumental in facilitating research that is being published in high-impact journals, due to both its richness and reliability.

However, accessing data from DST can be a lengthy and complicated process, due to the stringent requirements involved. Researchers must navigate a detailed application process that includes securing affiliation with an authorised Danish institution, submitting detailed project proposals and adhering to strict data protection regulations. These requirements might present considerable barriers to access for international researchers or for those from smaller institutions.

The complexity of the data itself poses additional challenges. The vast array of datasets, each with unique parameters and formats, demands a significant level of expertise and understanding that can be time-consuming and resource-intensive to develop. Moreover, while information on registers is increasingly available in English, substantial progress still needs to be made in providing bilingual documentation.

In spite the room for improvement outlined above, Denmark has established a commendable research environment that provides rich, reliable data and continuously strives for enhancements. This proactive approach to data management and accessibility ensures that Denmark remains at the forefront of research innovation, thereby supporting a wide range of scientific and policy-oriented studies.

References

Bobbio, E., & Bunzel, H. (2018). *The Danish matched employer-employee data.* Institut for Økonomi, Aarhus Universitet. Economics Working Papers No. 2018-03. https://pure.au.dk/portal/en/publications/the-danish-matched-employer-employee-data

DST. (2016). The Danish system for access to micro data B5-15. Agreement + Appendix 1 + Appendix 2 + Appendix 3. https://www.dst.dk/Site/Dst/SingleFiles/GetArchiveFile.aspx?fi=5452354440&fo=0&ext=israel2016

DST. (2022). Statistical program DST 2022. ISBN 978-87-501-2402-3. ISSN 2794-3119 www.dst.dk/publ/statistikprogramUK

DST. (2023a). The Data Warehouse for Business Statistics: New and improved access possibilities to microdata for research and analysis. https://www.dst.dk/-/media/Kontorer/13-Forskning-og-Metode/Data-Warehouse_digital_final2.pdf

DST. (2023b). Statistical program DST 2023 ISBN 978-87-501-2428-3 ISSN 2794-3119. www.dst.dk/publ/statistikprogramUK

DST. 2024. Statistical program DST 2024, ISBN 978-87-501-2447-4. ISSN 2794-3119. www.dst.dk/publ/statistikprogramUK

Emerek, R., Vejrup-Hansen, P., & Leth-S_rensen, S. (1991). Ida- en integreret database for arbejdsmarketforskning.

Folketinget. (2018). Lov om Danmarks Statistik LBK nr 610 af 30/05/2018. https://www.dst.dk/en/OmDS/strategi-og-kvalitet/kvalitet-for-statistikproduktion/lovgivning

Nordic Labour Journal, "Denmark leads the way on statistics using microdata". Article Dec. 17. 2018. http://www.nordiclabourjournal.org/i-fokus/in-focus-2018/theme-nordic-statistics/article.2018-12-17.2801676342

Timmermans, B. (2010). *The Danish integrated database for labor market research: Towards demystification for the English speaking audience.* DRUID, Copenhagen Business School, Department of Industrial Economics and Strategy/Aalborg University, Department of Business Studies, DRUID Working Papers. https://vbn.aau.dk/en/publications/the-danish-integrated-database-for-labor-market-research-towards

Maria Schlier is a Postdoctoral Fellow at University of Copenhagen. She obtained his PhD in Economics from Copenhagen University in 2022, a Masters degree from the Paris School of Economics and Bachelor degree from the Toulouse School of economics. Her areas of research includes corporate governance, leadership and applied microeconomics.

Morten Bennedsen is Niels Bohr Professor at the University of Copenhagen and André and Rosalie Hoffmann Chaired Professor of Family Enterprise at INSEAD and the Academic Director of the Wendel International Centre for Family Enterprise INSEAD. He obtained his PhD in Economics from Harvard University in 1998, a Master's degree with distinction from London School of Economics and a Bachelor Degree from the University of Copenhagen. His area of teaching includes family business governance, corporate governance, applied microeconomics and firm theory. Morten Bennedsen's main research area is the governance of family firms and other closely held corporations in a global context. His work has been published in top finance and economics journals, such as Journal of Finance, Journal of Financial Economics, Review of Financial Studies, Journal of Political Economics and Quarterly Journals of Economics and he is the author of the book "The Family Business Map: Assets and Roadblocks in Longterm Planning". Morten Bennedsen has won a number of research and teaching prizes including the biggest Danish individual research prize from the Ministry of Science.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 9
Estonian Administrative Data for Policy Evaluation



Andres Võrk, Kadri Rootalu, and Jaan Masso

Abstract Estonia's advanced digital infrastructure facilitates the use of administrative data for policy evaluation. Unique personal identification codes, population-wide registers, the X-Road data exchange system, and Statistics Estonia (acting as an intermediary) enable the secure and efficient access to registry data across key policy domains, including labour markets, business, health, and social insurance. Most registers date back to the late 1990s or to the early 2000s. The absence of a single access point requires researchers to navigate a fragmented approval process, one that involves multiple registry owners. While Statistics Estonia provides a centralised platform for data analysis, its computational limitations may restrict the use of modern machine-learning models. Estonia's administrative data system offers a strong foundation for evidence-based policymaking. However, improving data integration, computational capacity, and regulatory efficiency will be essential for enhancing the usability of administrative data for policy evaluation.

1 Introduction

This chapter provides an overview of how researchers and policy analysts might use Estonian registry data for policy evaluation. It focuses on personal-level data and matched employer-employee data, covering data sources typically required for policy evaluations in economics, social policy, or health policy.

First, we outline the ecosystem that supports Estonia's e-government and research infrastructure. This is followed by an overview of the legislative framework and

A. Võrk (✉) · J. Masso
University of Tartu, Tartu, Estonia
e-mail: andres.vork@ut.ee

J. Masso
e-mail: jaan.masso@ut.ee

K. Rootalu
Statistics Estonia, Tallinn, Estonia
e-mail: kadri.rootalu@stat.ee

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_9

procedures by which to access registry data. Next, we provide a brief overview of the main socio-economic registries that are available for policy evaluation and causal inference. Finally, we discuss options for merging registries, with a particular focus on the interface provided by Statistics Estonia.

2 Ecosystem for Using Administrative Data for Policy Evaluation

Estonia has a modern digital ecosystem that enables researchers and policy analysts alike to systematically use Estonian administrative data for ex-post and ex-ante policy evaluations. Regular evaluations are conducted by in-house analysts within public institutions as well as by researchers at universities and think tanks. The topics range from traditional economic questions, such as the optimal design of the tax-benefit system or the effects of subsidies on enterprise performance, to demographic issues, such as the determinants of fertility.

Estonia consistently ranks as a high performer in comparative measures of digitalisation (Solvak & Lauringson, 2024). The Organisation for Economic Cooperation and Development's Digital Government Index 2023 ranks the country sixth globally in the overall index and second in data-driven governance, which measures a government's progress in developing the necessary foundations to facilitate data access and sharing across the public sector.[1] The digital ecosystem supports the use of administrative data for policy evaluations and is built around several key components.

First, a unique personal identification code (*isikukood*), which is automatically assigned to every Estonian citizen at birth, enables individuals to be uniquely identified across different registries. Foreign nationals living in Estonia who have obtained a residence permit or right of residence are also assigned this code. Additionally, non-residents who need a personal identification code for administrative purposes (e.g. to open a bank account, register a company, or when interacting with Estonian authorities, such as Statistics Estonia) can apply for it. Individuals participating in Estonia's e-Residency programme[2] are also issued a personal identification code, thereby enabling them to access Estonia's digital services, such as starting a company.

This personal identification code is used in all administrative registries and allows for the unique identification of individuals and for the merging of data across different registries. From the data user's perspective, the code also helps to track who accesses registry data. For example, foreign researchers seeking access to pseudonymised Estonian administrative data, via Statistics Estonia's Remote Desktop Protocol, must first obtain a personal identification code and an ID card. This is now possible through the e-Residency programme.

Second, Estonia utilises population-wide registers. As a small country, most taxes and benefits are applied at the national level. Single registries contain all relevant

[1] OECD Digital Government Index, https://goingdigital.oecd.org/en/indicator/58.

[2] E-Residency of Estonia. https://www.e-resident.gov.ee/.

information for policy analysis, thereby eliminating the need to merge data from different regions. Still, some benefits and services, primarily means-tested, are implemented at the municipal level, with detailed data stored in local registries. However, even then the key information on local benefits and services is pooled into a central registry, but it is of lower quality due to variations in data entry practices.

Third, Estonia has developed the X-Road data exchange system that enables secure communication between various public and private information systems in Estonia.[3] Its primary purpose is to facilitate both the provision of and access to public and private e-services. As a by-product, it also allows public institutions to access data from different registries for policy evaluations. However, such access may require permission from the Estonian Data Protection Inspectorate.

Fourth, researchers from both Estonia and abroad can use Statistics Estonia as an intermediary that provides data pseudonymisation services and the linking of different administrative and survey data. Researchers can access data via secure VPN connections. However, Statistics Estonia's environment does not include datasets that are not used to produce national statistics, such as health information or genetic data. Other types of information can be integrated into the system if the necessary permissions are obtained from data owners, ethics committees and the Data Protection Inspectorate, with the exception of genetic data.

The Estonian government also promotes the use of open data. The Estonian open data portal provides an overview of all administrative open datasets that are not governed by the GDPR.[4] Annual reports (balance sheets and income statements) and tax data (quarterly data on payroll tax, sales and employees) on individual companies are particularly useful for policy evaluators. These datasets are published by the Business Register[5] and the Estonian Tax and Customs Board,[6] respectively. The latter, published on a quarterly basis, becomes available shortly after the end of the respective quarter.

3 Legislation and Procedures Governing Analysis of Personal and Business Data

The use of personal data for policy evaluation is, at the most general level, governed by the General Data Protection Regulation (GDPR). The Personal Data Protection Act[7] regulates the protection of natural persons' personal data, elaborating on the

[3] X-Road – interoperability services. https://e-estonia.com/solutions/interoperability-services/x-road/.

[4] Estonian Open Data Portal, https://andmed.eesti.ee/.

[5] Open data of the e-Business Register, https://avaandmed.ariregister.rik.ee/en.

[6] Estonian Tax and Customs Board, Statistics and open data, https://www.emta.ee/en/private-client/board-news-and-contacts/news-press-information-statistics/statistics-and-open-data.

[7] Personal Data Protection Act, passed 12.12.2018, published 04.01.2019, last changed 16.03.2026, https://www.riigiteataja.ee/en/akt/519032026002.

EU's GDPR and addressing data processing by law enforcement authorities. The act also defines the conditions under which personal data can be used for scientific research. It stipulates that personal data may be processed, without the data subject's consent, for the purposes of scientific and historical research, as well as official statistics, particularly in a pseudonymised format or in a format that provides an equivalent level of protection.

The Personal Data Protection Act (simply "act" in the following) states that the processing of personal data for the purposes of scientific and historical research or official statistics, without the data subject's consent and in a format that enables their identification, is permitted only if the following three conditions are met:

(1) the purposes of data processing cannot be achieved after the removal of identifying data, or it would be unreasonably difficult to achieve these purposes;
(2) there is overriding public interest in the processing of the data, as determined by those conducting scientific and historical research or compiling official statistics;
(3) the processing does not alter the scope of the data subject's obligations based on the processed personal data, nor does it affect the data subject's rights in any other way.

Obtaining permission to conduct a study is required in cases specified by the act. Such permission must be sought either from an ethics committee in the relevant field of research or from the Data Protection Inspectorate[8] or, in certain cases, from both. When a researcher applies for access to personal data, they must demonstrate that all three aforementioned conditions are met.

The act stipulates that if research is based on special categories of personal data (as defined by the GDPR), then the ethics committee in the relevant research field must first verify compliance with the terms and conditions. Therefore, permission from the ethics committee is required whenever data about ethnicity, trade union membership, religious affiliation, or health data—or even disposable income data from registers if it is obtained through the processing of monetary healthcare benefits. There are several ethics committees in Estonia whose agreement researchers need if they work with special categories of personal data or with clinical data. They either fall under universities, Estonian Research Council, or under the Ministry of Social Affairs.[9]

Scientific research also includes analyses and studies conducted by government agencies for the purposes of policy development. Agencies have the right to query the databases of other agencies and process the personal data received in order to prepare these analyses. However, in such cases, the Estonian Data Protection Inspectorate must give its consent before the specified processing of personal data begins. If the objectives of studies conducted for policy development and the scope of personal data processing derive directly from legislation, then consent from the Estonian Data Protection Inspectorate is not required.

[8] Data Protection Inspectorate, https://www.aki.ee/en.

[9] Ethics committees under the jurisdiction of the Ministry of Social Affairs (in Estonian) https://www.sm.ee/uudised-ja-pressiinfo/uuringud-ja-statistika/valitsemisala-eetikakomiteed.

Permission from the Data Protection Inspectorate or from an ethics committee is not required if personal data is processed in scientific research with the individual's consent. In such cases, general data protection regulations apply to the processing of personal data. Permission is also not required if the scientific research is conducted with anonymous data, given that such data is not considered personal data under the law. However, permission from the ethics committee and from the Data Protection Inspectorate may be required to anonymise personal data before it is provided to a researcher, for example, by the registry owner.

Permission from the Data Protection Inspectorate must be obtained if a government institution conducts a policy development study, or procures a study from a research institute, and the data repository is not managed by itself, but by another data controller or processor. Permission from the ethics committee must be obtained before applying to the Data Protection Inspectorate for policy development studies involving special categories of personal data.

The use of firm-level data by researchers is not regulated, provided that the data is publicly available, such as companies' annual reports, information on subsidy and grant receipts, or open data on taxes paid, as published by the Estonian Tax and Customs Board. However, when firm-level data is collected by Statistics Estonia for statistical purposes, or by the Estonian Tax and Customs Board for taxation purposes, such data are generally made available to researchers or policy analysts only via VPN connection to Statistics Estonia. Standard confidentiality checks are applied when taking aggregate data out for research publications.

One example of these datasets is the international goods trade dataset (available since 1995), which includes values and quantities of exports and imports by company, destination (or source) countries and 8-digit Combined Nomenclature product codes (see, e.g., Benkovskis et al., 2020).

The use of data via Statistics Estonia is regulated by the Official Statistics Act.[10] This act explicitly states that data collected from administrative records and databases may also be transmitted to researchers for scientific purposes. If a research institute wants to use registry data for research within Statistics Estonia, and all data from different registries is already in the possession of Statistics Estonia for generating national statistics, then permission from the Data Protection Inspectorate is not required. However, if special category data is included, then approval from an ethics committee is still necessary.

[10] Official Statistics Act. Passed 10.06.2010, RT I 2010, 41, 241, entry into force 01.08.2010, last modified 21.03.2022, https://www.riigiteataja.ee/akt/111032022002.

4 Data Coverage

Most policy evaluation studies rely on administrative data collected by government authorities, and made available retrospectively, as part of their routine operations. In many cases—particularly regarding firm-level data—Estonian datasets offer reasonably long-time coverage, now approaching thirty years in some cases.

Administrative data is often merged with data from regular surveys conducted by Statistics Estonia, thereby enabling the combination of different data categories (e.g. self-assessed health and income, or enterprise balance sheets with innovation survey data). In the following section, we provide a brief overview of existing administrative data sources frequently used for policy evaluations, either by public institutions or by research organisations. We cover data sources relevant to policy evaluations in economics, social policy and health policy. The time period in the tables below refers to data available to researchers, according to the best of our knowledge.

The administration system for the state information systems (RIHA), managed by the State Information System Authority, provides descriptions of information systems and data.[11] However, these descriptions are primarily intended for information system developers and are not as suitable for researchers. Below we give references to both RIHA database and provide links to additional information sources.

4.1 Labour Market Data

The primary source of administrative employment data comes from the registries of the Estonian Tax and Customs Board. Its Employment Registry (*Töötamise register*) consolidates all of the information related to employment since 2014. All employers are required to register employment. This registry serves as the basis for determining eligibility for social benefits associated with employment (e.g. health insurance, unemployment benefits, pension contributions, etc.).

One drawback is that the register does not include exact work hours; it just provides information about whether the job is full-time or part-time. Since 2019, it has also included information on employees' occupations (it took nearly ten years for Statistics Estonia to begin collecting occupational data on a longitudinal basis).

The Tax and Customs Board also maintains records on monthly labour taxes (the Income and Social Tax Return—*TSD Tulu- ja sotsiaalmaksu, kohustusliku kogumispensioni ja töötuskindlustusmakse deklaratsioon*) reported by employers, as well as on annual income tax return data (*Tuludeklaratsioon*).

The registries of the Tax and Customs Board are available to researchers only through Statistics Estonia's Remote Desktop Protocol where the data is pseudonymised. Since the Employment Registry and monthly earnings data have

[11] Riigi Infosüsteemi Haldussüsteem, Infosüsteemid, https://www.riha.ee/Infos%C3%BCsteemid.

Table 9.1 Main labour market administrative data for policy evaluation

Data source	Owner	Main variables	Time period
Employment register[12] (*Töötamise register - TÖR*)	The Estonian Tax and Customs Board	Individual's and employer's code, type of employment, dates, ISCO code (since 2019), workload (optionally), location of work	2014 -
Annual Income Tax Return Data[13]	The Estonian Tax and Customs Board	Individual's yearly income by income source	1999 -
Monthly income and social tax return data (TSD)[14]	The Estonian Tax and Customs Board	Individual's and employer's code, monthly gross earnings, taxable benefits, taxes paid	2006 - (in Statistics Estonia)
The Unemployment Insurance Fund Database[15]	The Estonian Unemployment Insurance Fund	Data on registered unemployment spells, all labour market measures both for the unemployed and employed, unemployment benefits, work ability assessment, disability status and related benefits	Varies
Income database[16]	Statistics Estonia	Annual data for all income sources used for the calculation of registry-based poverty rates	2017-

Source Compiled by the authors

been widely used by Estonian researchers (see, e.g., Masso et al., 2022, for an analysis of the gender pay gap), this data is already available on Statistics Estonia's servers and often requires only limited additional processing by Statistics Estonia.

All registries include pseudo-identification codes of enterprises and individuals and this allows researchers to merge employer-employee data and to add background characteristics from other data sources, such as the population register, education register and so on (Table 9.1).

[12] A formal overview of the register is offered here: https://www.emta.ee/en/business-client/registration-business/employment-register, Statistics Estonia can provide an overview of which variables are available for researchers.

[13] Tax return forms are available at https://www.emta.ee/en/private-client/taxes-and-payment/declaration-income/tax-return-forms-and-instructions, Statistics Estonia can provide an overview of which variables are available for researchers.

[14] TSD return form is available https://www.emta.ee/en/business-client/taxes-and-payment/income-and-social-taxes/submission-declaration-form-tsd, Statistics Estonia can provide an overview of which variables are available for researchers.

[15] Statutes of the Unemployment Insurance Fund Database (*Töötukassa andmekogu põhimäärus*, in Estonian) available at https://www.riigiteataja.ee/akt/110052024019, the EUIF can provide an overview of which variables are available for researchers.

[16] Statistics Estonia can provide an overview of which variables are available for researchers.

All registered unemployment data is collected by the Estonian Unemployment Insurance Fund (*Eesti Töötukassa*—EUIF). The EUIF database contains comprehensive information on unemployment spells, active labour market services and unemployment benefits. It also includes previous labour market history (derived from the Employment Registry) and labour earnings (derived from monthly tax returns) using the X-road interface. It also includes information on work ability assessments and benefits related to reduced work ability.

This is a highly detailed and valuable dataset for labour economists. However, as of the beginning of 2025, it has not been widely used by researchers. Consequently, only limited information is available via Statistics Estonia, such as data on unemployment spells and some basic details on labour market services and benefits currently used in producing official statistics. If more detailed data is needed for merging with other registries, then the standard procedure involving ethics committees and, if required, the Data Protection Inspectorate—outlined in Sect. 2—must be followed.

The EUIF uses its own database for regular labour market evaluations. It has developed a counterfactual impact evaluation tool called MALLE that employs a combination of exact matching and propensity score matching to evaluate the effect of labour market services on subsequent re-employment probability and earnings. Published applications include analyses of the effects of vocational training and work practice (Eesti Töötukassa, 2024a, 2024b).

4.2 Business Data

Business administrative data for causal analysis primarily consists of data on annual reports, taxes and foreign trade. This data is often linked within Statistics Estonia to employee register data or various survey data conducted by Statistics Estonia (such as innovation surveys, research and development data). Such linking has been implemented within specific research projects, as well as within the Estonian Research Infrastructures Roadmap project "Infotechnological Mobility Observatory (IMO)".[17] Analogously to the individual-level data, all legal entities also have a unique identification code (*registrikood*), which can be used to merge the various firm-level datasets. Statistics Estonia has also cleaned business registry data and linked it to basic annual statistics (e.g. employment and turnover) to its Statistical Business Registry (*Majandusüksuste statistiline register*) and this can also be used by researchers (Table 9.2).

A rather unique dataset that contains information on inter-company transactions, i.e. flows between companies that give supplier and customer linkages, is the appendixes of the VAT declarations. In previous studies, this dataset was used to study the regional economic structures of Estonia (Martyanova, 2019), knowledge transfer from multinational companies to domestic companies (Masso & Vahter, 2023) and the diffusion of green and digital technologies (OECD, 2023).

[17] https://imo.ut.ee/en/.

Table 9.2 Main firm-level administrative data

Data source	Owner	Main variables	Time period
Business Register Data[18]	Business register	Establishment of firm, owners, board members, industry	1995-
Annual reports of firms[19]	Business register	Balance sheets and income statement data	1995-
Statistical business registry (MSR—majandusüksuste statistiline register)[20]	Statistics Estonia	Cleaned business register data used by Statistics Estonia	1995-
EKOMAR survey of structural business statistics[21]	Statistics Estonia	Balance sheets and income statement data	1995-
Foreign trade of goods data[22]	Estonian Tax and Customs Board	Detailed level data on the export and import of goods at the firm level	1995-
Foreign trade of services data[23]	Bank of Estonia	Detailed level data on the export and import of services at the firm level	2006-
Outward FDI data[24]	Bank of Estonia	The sum of investments by the destination countries	1995-
Corporate VAT declarations (KMD) data[25]	Estonian Tax and Customs Board	Information on VAT declarations Flows between companies (invoices larger than 1,000 EUR)	2014-

Source Compiled by the authors

It should be borne in mind that conducting network analysis at Statistics Estonia is restricted due to confidentiality rules, which prohibit the release of data on individual statistical units (i.e. individuals or companies) outside the institution. This means that while visualisation of individual data points is not allowed, aggregate network characteristics or the results of statistical models are permitted.

[18] https://ariregister.rik.ee/eng.

[19] https://avaandmed.ariregister.rik.ee/en/downloading-open-data.

[20] Statistics Estonia can provide an overview of which variables are available for researchers.

[21] https://stat.ee/en/submit-data/about-data-submission/enterprise-surveys/ekomar, Statistics Estonia can provide an overview of which variables are available for researchers.

[22] https://www.stat.ee/en/find-statistics/methodology-and-quality/esms-metadata/22303, Statistics Estonia can provide overview which variables are available for researchers.

[23] https://stat.ee/en/find-statistics/methodology-and-quality/esms-metadata/22304, Statistics Estonia can provide an overview of which variables are available for researchers.

[24] More detailed information is available from Eesti Pank, https://www.eestipank.ee/en/statistics.

[25] https://www.emta.ee/en/business-client/taxes-and-payment/value-added-tax/filing-vat-returns-and-reports, Statistics Estonia can provide an overview of which variables are available for researchers.

Table 9.3 Main administrative data sources on social benefits and services

Data source	Owner	Main variables	Time period
Social Insurance Board register data[26]	Social Insurance Board	State pension scheme: pension rights, contributions, pension size, dates, etc. Family benefits Disability benefits Disability related services, etc.	1999-
STAR[27]	Social Insurance Board	Social services and local benefits by local municipalities at individual or household level	2010-
Funded pension schemes[28]	Owned by Ministry of Finance, managed by Pensionikeskus AS	Contributions to and payments from the funded pension schemes at individual level Value of pension assets by pension funds	2002-

Source Compiled by the authors

Statistics Estonia is responsible for collecting various firm-level data via surveys, often coordinated by Eurostat. These data are also made available to researchers via Remote Desktop Protocol.

4.3 Social Insurance and Assistance Data

The largest administrative data on social insurance benefits is available from the Estonian Social Insurance Board (*Sotsiaalkindlustusamet*), which is responsible for administering pensions, family benefits, disability-related benefits, and funding certain rehabilitation services and devices.

This rich dataset is often merged with other registries in order to analyse the effects of changes in the social insurance system (Table 9.3). For example, it has been used to study the impact of parental benefits on fertility behaviour (Abdullayev & Puur, 2024; Puur et al., 2023) and to analyse the effects of pension system reforms (Piirits & Võrk, 2019; Soosaar et al., 2021).

The Estonian pension system also includes funded pension schemes. Data on these schemes is available from Pensionikeskus. Researchers must combine individual-level data on state pensions from the Social Insurance Board and Pensionikeskus in order to conduct ex-ante or ex-post evaluations of the reforms affecting total old-age pension.

[26] https://www.riha.ee/Infos%C3%BCsteemid/Vaata/skais (description in Estonian).

[27] https://www.riha.ee/Infos%C3%BCsteemid/Vaata/star (description in Estonian).

[28] https://www.riha.ee/Infos%C3%BCsteemid/Vaata/kpr (description in Estonian), https://nasdaqcsd.com/estonia/en/esis-id/pension-services/.

Finally, administrative data on services and benefits provided by local municipalities is collected within the information system STAR, which is also managed by the Social Insurance Board. This register must be included when focusing on means-tested benefits.

4.4 Administrative Health Care Data

The main sources of health data are the Estonian Health Information System and the Estonian Health Insurance Fund. The Estonian Health Information System is a centralised national e-health platform designed to securely manage, store and exchange health-related information for individuals, healthcare providers and public health authorities. It stores electronic health records for all residents. All Estonian healthcare providers have an obligation to submit information to the registry. The use of the system is challenging because it contains free-text data. Estonian researchers have worked on standardising the data to improve its usability for research (see, Talvik et al., 2025).

Alternatively, researchers have used the registries of the Estonian Health Insurance Fund to define the health of individuals or healthcare utilisation. The fund has provided universal public health insurance since 2001 and this covers over 95% of the Estonian population. The fund maintains a comprehensive healthcare and prescription pharmaceutical database that forms the basis for the payments made to healthcare providers. The main drawback of the EHIF database is that it does not cover healthcare services paid directly by patients, their employers or through voluntary health insurance. The share of out-of-pocket payments is particularly high in dental care and ambulatory special care visits, thereby leading to an underestimation of the healthcare utilisation or diagnoses in the EHIF data (Kasekamp et al., 2023).

It should be borne in mind that Statistics Estonia does not use health registries for regular statistics, and as of the beginning of 2025, researchers have not used Statistics Estonia's environment for health data analysis. Instead, they have requested data directly from respective data owners. Approval from an ethics committee is required in order to access health data. Specifically, approval from the Research Ethics Committee of the Estonian Research Council is required in order to access Estonian Health Information System data for research purposes (Table 9.4).

Some administrative health registries are available through the National Institute of Health Development, a national research and development institution responsible for public health research, disease prevention programmes and health promotion activities.

The Estonian Biobank (*Eesti Geenivaramu*), managed by the University of Tartu's Institute of Genomics, is a large-scale population-based biobank established in order to support genetic research and promote personalised medicine (Milani et al., 2025). The biobank contains data from over 200,000 gene donors, representing more than 20% of Estonia's adult population. Genetic data is primarily used for health-related studies and personalised prevention, but they have also been utilised to examine the

Table 9.4 Main administrative data sources on health

Data source	Owner	Main variables	Time period
Estonian Health Information System[29]	Ministry of Social Affairs	Electronic health records	2008-
Health insurance data[30]	Estonian Health Insurance Fund	All data necessary for reimbursement of services and medications covered by health insurance	2005-
Specific registers[31]	National Institute of Health Development	E.g. cancer register, cancer screening register, causes of death register, medical birth register	Varies
Estonian Biobank[32]	University of Tartu	Genomics, phenotypes	2002-

Source Compiled by the authors

relationship between genes and socio-economic outcomes (see, for example, Kweon et al., 2025; Rimfeld et al., 2018).

4.5 Other Population Socio-Economic Data

Field-specific administrative data are usually linked to general background information about individuals, such as education, mother tongue, place of residence, number and age of children, and household composition. This information is typically collected through population censuses. The most recent population census was register based, relying on various administrative sources. Census data is also available through Statistics Estonia.

The Population Register is a database that consolidates primary personal data on individuals. It links children to their parents and records addresses. The register is both managed and was developed by the Ministry of the Interior and is used to carry out public tasks by the state, local governments and other legal entities (Table 9.5).

Education information is available as self-reported data in the census or population register. Data from the Estonian Education Information System can be used for more recent years (from 2004 forward). It includes information about all studies of

[29] https://www.riha.ee/Infos%C3%BCsteemid/Vaata/digilugu (description in Estonian), for data queries contact Health and Welfare Information Systems Centre (TEHIK), https://www.tehik.ee/en/contact.

[30] https://www.riha.ee/Infos%C3%BCsteemid/Vaata/kirst (description in Estonian), for data queries, see https://www.tervisekassa.ee/andmeparingud.

[31] The National Institute of Health Development, "Statistics and registers", https://www.tai.ee/en/statistics-and-registers.

[32] https://genomics.ut.ee/en/content/estonian-biobank.

Table 9.5 Main administrative data sources on background statistics

Data source	Owner	Main variables	Time period
Population and Housing Census data[33]	Statistics Estonia	Main socio-economic variables, information on households, place of living, etc.	1989, 2000, 2011, 2021
Population register[34]	Ministry of the Interior	Marital status, spouse, children, parents, place of living, self-reported ethnicity, mother tongue, education level	2002-
Statistical population register[35]	Statistics Estonia	The statistical registry contains population data for statistical purposes, combining various registers in addition to the population registry	2010-
Estonian Education Information System[36]	Ministry of Education and Research	Data on educational attainment, learning and graduation episodes. The register contains data from primary and extracurricular education to higher education; it also includes information about educators	2004-

Source Compiled by the authors

individual's, including field of study and curriculum group; these are categorised according to the ISCED 2011 classification and updated annually.

Statistics Estonia also hosts specific merged datasets for simulation models. Registry-based EUROMOD microsimulation models enable detailed ex-ante simulations of the Estonian tax-benefit system and is widely used by researchers (Laurimäe et al., 2022) as well as policy analysts at the Ministry of Finance and the Ministry of Social Affairs. In 2024–2025, a dynamic microsimulation model using population-level registry data was being developed and will be made available to researchers once complete.

Additionally, Statistics Estonia provides researchers access to data from the OECD Survey of Adult Skills (PIAAC), which is merged with various Estonian registry data.

[33] https://rahvaloendus.ee/en/census-2021/contacts, Statistics Estonia can provide an overview of which variables are available for researchers,

[34] https://www.riha.ee/Infos%C3%BCsteemid/Vaata/rr (description in Estonian).

[35] Statistics Estonia can provide an overview of which variables are available for researchers,

[36] https://www.riha.ee/Infos%C3%BCsteemid/Vaata/EHIS (description in Estonian); public data https://www.ehis.ee/; Statistics Estonia can provide an overview of which variables are available for researchers.

5 Accessing the Data

5.1 Statistics Estonia Remote Connection

A convenient and commonly used approach for policy evaluations with registry data in Estonia is to use Statistics Estonia as an intermediary for data analysis, given that it already houses data from many registries for its routine production of national statistics. All details about applying for data from Statistics Estonia and using confidential data are available on the Statistics Estonia web page.[37]

Agreements with Statistics Estonia for access to confidential data for scientific purposes can only be made with a legal entity, not through an individual. In the agreement, the legal entity must specify the projects for which the data will be used. Typically, an organisation will submit an application to Statistics Estonia, specifying the list of requested data. Statistics Estonia then prepares a pseudonymised dataset, accessible via the data processing environment, along with metadata. If data is not available in Statistics Estonia, then researchers can request registries to prepare personal-level data and send it to Statistics Estonia. The data is transmitted, in an encrypted form, or is submitted through a data gateway. Statistics Estonia then pseudonymises the datasets.

The prepared data can be accessed remotely over VPN via Statistics Estonia's secure workstation, which requires dual authentication (including an Estonian ID card). This is also possible from abroad through the e-residency programme, as explained in Sect. 3. The Statistics Estonia server has Stata, R, Python, SPSS, SQL and Open Office tools for data analysis.

Statistics Estonia always signs a confidentiality agreement with the organisation and with all authorised users from the organisation. The processing of data can be done by the research organisation's authorised representatives.

Results of the analysis can be exported from the secure environment only in an aggregated format, ensuring that individual data subjects cannot be identified. Users may export aggregated data only after receiving Statistics Estonia's approval. Statistics Estonia applies statistical disclosure control and provides researchers with detailed instructions on what qualifies as an acceptable scientific output. If necessary, the Dutch Statistical Office's τ-ARGUS software is used to verify the non-identifiability of specific units (e.g. enterprises or individuals).[38]

The default Statistics Estonia server for researchers has physical limits (memory, speed, storage), making it unsuitable for large machine-learning models. Therefore, researchers primarily use structured data and classical econometric tools in the Statistics Estonia server. Some administrative datasets, for example, Estonian Health Information System data, require substantial resources and therefore Estonian researchers

[37] Statistics Estonia. *Guidelines for using confidential data* https://www.stat.ee/en/guidelines-for-using-confidential-data; *Use of confidential data for scientific purposes.* https://stat.ee/en/find-statistics/request-statistics/use-confidential-data-scientific-purposes.

[38] https://research.cbs.nl/casc/tau.htm.

use it on their own servers. However, additional memory and computing power can be ordered from Statistics Estonia, provided the researcher has the necessary funding, for large-scale data analysis needs.

Some metadata-related information needs to be obtained from the original source because Statistics Estonia does not manage administrative registers itself. The data processed at Statistics Estonia already includes metadata in English (though not always in 100% of cases). Additionally, survey questionnaires, which are often linked to register data, are typically translated into English.

Although there is no single point of access to register data, the cumulative use of Statistics Estonia's services by researchers has steadily expanded both the availability and integration of these data sources. Since Statistics Estonia does not have access to all registers for its routine statistical production, the more data researchers use through Statistics Estonia, the greater the volume of register data, including metadata, that becomes available within its environment.

Statistics Estonia is currently developing a new data exchange service called ERIKA, a Trusted Research Environment. In future, ERIKA is expected to streamline processes for accessing data descriptions, ordering data, managing and modifying contracts, and verifying the confidentiality of research outputs.

Research conducted by scientists and policy analysts is considered non-programmatic statistical work for Statistics Estonia, meaning researchers or policy analysts must finance this work themselves. The pricing for Statistics Estonia's services is publicly available and consists of a fixed fee and a data preparation fee that is calculated based on working hours needed for the task. In 2025, the fixed cost for entering into an agreement was 530 euros. The price of contract work depends on the time required and on the task's complexity. No additional costs are incurred in cases for which the data are already available in the format that is suitable for the researcher. The preparation fee may easily reach 2,000 euros when the number of datasets to be prepared is larger and the researchers also need the data in pre-processed format (for example, different data sources linked over a longer time period). Simpler tasks like preparing 1–2 data sources from more recent years with minimal preprocessing would not exceed 150 euros. As a rule, a cost estimate for contract work is prepared within five working days, and the work is carried out only after the customer accepts the cost estimate.

5.2 Other Options

Accessing personal-level data for the purposes of policy evaluation can also be done directly by requesting data, with personal identification codes, from registries and merging the data independently. Researchers must be familiar with the data structure in the registries in order to make such queries. Typically, this requires the research institution to sign data confidentiality agreements with the registry owner. Individual-level data is not given to individual researchers, but institutions, and all data protection requirements set by the user organisation must be met. Usually, this means that a

designated responsible person within the research project will pseudonymise the data and will make it available to project researchers, for example, on a secure server.

However, it is important to note that data owners are not obligated to provide researchers with individual-level data. For instance, the Estonian Tax and Customs Board has made its data available through Statistics Estonia and does not provide direct access to researchers. Conversely, Estonian researchers can apply for health data directly since health registers are not used by Statistics Estonia.

A few studies in Estonia have used the proprietary tool Sharemind to merge registry data.[39] Sharemind is a privacy-preserving technology designed to enable secure analysis and linkage of data from multiple sources without compromising the confidentiality of sensitive information. Developed by the Estonian company Cybernetica, Sharemind uses advanced cryptographic techniques to facilitate secure computations. For example, when tax data were not available to researchers via Statistics Estonia, Sharemind was used to link the education registry and tax data (Bogdanov et al., 2016; Cybernetica, 2017). Policy evaluators and researchers themselves no longer use it, as more flexible ways of merging registry data are available through Statistics Estonia. Sharemind is also considered to be one possible tool for secure multiparty computation by Eurostat within the JOCONDE project.[40]

6 Summary

Estonia has developed a highly advanced digital ecosystem that enables comprehensive policy evaluation by using administrative data. The system is built on several key components, including unique personal identification codes, population-wide registers, the X-Road data exchange platform and Statistics Estonia as a central data intermediary. These components all facilitate secure and efficient data access, making Estonia a leader in digital governance and e-government services.

One of the Estonian system's greatest strengths is its extensive coverage of administrative data across key policy domains. Researchers have access to a wide range of datasets in areas as diverse as labour markets, health, social insurance, business and education. This wealth of information allows for regular policy analysis when different registers are linked. Centralised access through Statistics Estonia further strengthens the system by providing researchers with pseudonymised, and merged registry, data through secure workstations that are accessible via a VPN-based Remote Desktop Protocol.

Another of the Estonian administrative data system's notable strengths is its well-developed regulatory framework for data protection and data sharing. The country aligns with the General Data Protection Regulation (GDPR). The Personal Data

[39] Cybernetica, Sharemind—Privacy Enhancing Technology for Data-Driven Business. https://sharemind.cyber.ee/.

[40] Eurostat, Joint On-demand Computation with No Data Exchange, https://cros.ec.europa.eu/joconde.

Protection Act ensures strict privacy controls and statistical disclosure measures that safeguard individual confidentiality. The Official Statistics Act regulates how data collected from administrative records and databases by Statistics Estonia may be transmitted to researchers for scientific purposes. Estonia has fostered a growing culture of open data, with government agencies making a range of administrative datasets publicly available, such as tax and business statistics, through open-access portals.

Several challenges remain, despite the aforementioned strengths. One of the main limitations is the absence of a single point of access to registry data. Although the administration system for the state's information systems provides descriptions of information systems and data, it is not suitable for researchers. Researchers often need to request metadata and data separately from different registry owners, thereby making the process time-consuming and fragmented. This problem should be solved by streamlining access to metadata and research data, with the upcoming ERIKA Trusted Research Environment at Statistics Estonia, thereby simplifying the process for researchers.

While Statistics Estonia provides access to many datasets, certain types of data remain outside of its scope. Health data or detailed labour market policy data is not used by Statistics Estonia for national statistics, meaning that researchers must apply to the data owners directly in order to obtain access. At the same time, tax data is only available via Statistics Estonia, while direct access to these records is not permitted. The approval process for accessing sensitive data can also be administratively burdensome. Researchers must secure approvals from ethics committees and, in some cases, the Data Protection Inspectorate, making the process lengthy and complex.

In conclusion, Estonia's digital infrastructure provides a strong foundation for data-driven policy evaluation. The country has successfully built a system that facilitates both secure and efficient access to a broad range of administrative data, making it an attractive model for evidence-based policymaking. However, gaps remain in data accessibility, computing resources and the regulatory approval process. Strengthening data integration and streamlining the procedures for accessing sensitive datasets will be necessary steps in further enhancing Estonia's research and policy evaluation framework.

References

Abdullayev, S., & Puur, A. (2024). Varying responses to the introduction of earnings-related benefits: a study of 2004 parental leave reform in Estonia. *Journal of Population Research, 41* (4), ARTN 22. https://doi.org/10.1007/s12546-024-09342-1.

Benkovskis, K., Masso, J., Tkacevs, O., Vahter, P., & Yashiro, N. (2020). Export and productivity in global value chains: Comparative evidence from Latvia and Estonia. *Review of World Economics, 156*, 557–577. https://doi.org/10.1007/s10290-019-00371-0

Bogdanov, D., Kamm, L., Kubo, B., Rebane, R., Sokk, V., & Talviste, R. (2016). Students and taxes: A Privacy-preserving study using secure computation. *Proceedings on Privacy Enhancing Technologies, 3*, 117–135. https://doi.org/10.1515/popets-2016-0019

Cybernetica. (2017). National special education data analysed securely. https://cyber.ee/resources/news/national-special-education-data-analysed-securely/

Eesti Töötukassa. (2024a). Teenuste mõjuhindamise metoodika (*Methodology of evaluation of labour market services*). Report. https://www.tootukassa.ee/web/sites/default/files/2024-11/Teenuste%20m%C3%B5juhindamise%20metoodika_nov2024.pdf

Eesti Töötukassa. (2024b). Ametikoolituste mõjuhindamise tulemused (*Results of the evaluation of vocational training programmes*). Report. https://www.tootukassa.ee/web/sites/default/files/2024-01/Ametikoolituste%20m%C3%B5juhindamise%20tulemused.pdf

Kasekamp, K., Habicht, T., Võrk, A., Köhler, K., Reinap, M., Kahur, K., Laarmann, H., & Litvinova, Y. (2023). Estonia: Health system review. *Health Systems in Transition, 25*(5), 1–236.

Kweon, H., Burik, C. A. P., Ning, Y., et al. (2025). Associations between common genetic variants and income provide insights about the socio-economic health gradient. *Nature Human Behavior*. https://doi.org/10.1038/s41562-024-02080-7

Laurimäe, M., Paas, T., & Paulus, A. (2022). The effect of COVID-19 and the wage compensation measure on income-related gender disparities. *Baltic Journal of Economics, 22*(2), 146–166. https://doi.org/10.1080/1406099X.2022.2149976

Martyanova, I. (2019). *Analysis of regional economic network structures based on Estonian VAT transaction*. University of Tartu. http://hdl.handle.net/10062/64588

Masso, J., & Vahter, P. (2023). Joining and exiting the value chain of foreign multinationals and performance of their local suppliers: Evidence from interfirm transaction data. *Industrial and Corporate Change, 32*(4), 848–875.

Masso, J., Meriküll, J., & Vahter, P. (2022). The role of firms in the gender wage gap. *Journal of Comparative Economics, 50*(2), 454–473.

Milani, L., Alver, M., Laur, S., et al. (2025). The Estonian Biobank's journey from biobanking to personalized medicine. *Nature Communications, 16*, 3270. https://doi.org/10.1038/s41467-025-58465-3

OECD. (2023). *The market microstructure of industrial ecosystems in the digital and green transitions: How can production networks help design better industrial policies evidence from Estonia.* Mimeo.

Piirits, M., & Võrk, A. (2019). The effects on intra-generational inequality of introducing a funded pension scheme: A microsimulation analysis for Estonia. *International Social Security Review, 72*, 33–57. https://doi.org/10.1111/issr.12194

Puur, A., Abdullayev, S., Klesment, M., & Gortfelder, M. (2023). Parental leave and fertility: Individual-level responses in the tempo and quantum of second and third births. *European Journal of Population/Revue européenne de Démographie, 39*(1). https://doi.org/10.1007/s10680-023-09669-0

Rimfeld, K., Krapohl, E., Trzaskowski, M., Coleman, J. R. I., Selzam, S., Dale, P. S., Esko, T., Metspalu, A., & Plomin, R. (2018). Genetic influence on social outcomes during and after the Soviet era in Estonia. *Nature Human Behaviour, 2*(4), 269–275. https://doi.org/10.1038/s41562-018-0332-5

Solvak, M., & Lauringson, A. (2024, May 5). A case study of the public sector digital ecosystem in Estonia. *Computer 57*, 44–49. https://doi.org/10.1109/MC.2024.3375866

Soosaar, O., Puur, A., & Leppik, L. (2021). Does raising the pension age prolong working life? Evidence from pension age reform in Estonia. *Journal of Pension Economics and Finance, 20*(2), 317–335. https://doi.org/10.1017/S1474747220000244

Talvik, H.-A., Oja, M., Tamm, S., Mooses, K., Särg, D., Lõo, M., Siimon, Õ. R., Šuvalov, H., Kolde, R., Vilo, J., Reisberg, S., & Laur, S. (2025). Repeatable process for extracting health data from HL7 CDA documents. *Journal of Biomedical Informatics, 161*, Article 104765. https://doi.org/10.1016/j.jbi.2024.104765

Andres Võrk is a lecturer and analyst at the University of Tartu, School of Economics and Business Administration, Estonia. He teaches econometrics and policy evaluation methods. His research areas include labour and tax policy analysis, social security systems, and health care financing.

Kadri Rootalu is a team leader of the Population and Education service team at Statistics Estonia. She received her Ph.D. in sociology from the University of Tartu in 2018. Her primary research interests include education, labour market dynamics, family studies, and methodological innovation in the field of official statistics.

Jaan Masso is an Associate Professor at the University of Tartu, School of Economics and Business Administration, Estonia. His research areas include labour economics, foreign direct investments, innovations, international trade and research policy.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 10
Accessing Administrative Data for Research: Lessons from Finland



Stefano Lombardi

Abstract This chapter describes the rules and practices that govern access to administrative registers from Statistics Finland, the most common way to access Finnish socioeconomic register data. Overall, both the high data quality and the ability to integrate external microdata sources make the system exceptionally comprehensive and versatile. By describing the data access rules, this chapter also provides an assessment of the equality of access to the administrative data. Data access appears to be well developed and equal: it is centred around the national statistical authority and there are clear rules about how to access the data remotely; user permits applications are assessed, based on data protection rules, by data protection experts. Foreign-based researchers who are interested in remotely accessing Finnish registers can do so upon starting a collaboration agreement with a recognised Finnish institution. On the other hand, a potential source of inequality of access to the data stems from the increasing data costs that might potentially prevent younger and less well-established researchers from opening new projects.

1 Overview

This chapter has two main objectives. First, it aims to provide detailed information about how to get access to Finnish administrative registers for research purposes. Second, it aims to provide a description of the data protection rules that are in place at present, as well as an assessment of how equal the system is in terms of allowing researchers to access the data. As such, the chapter has been written in such a way as to keep both researchers and policymakers in mind. Researchers will be able to understand what kind of register-based data can be accessed in Finland; policymakers in Finland and elsewhere will get a practical sense of how rules have been implemented to protect the data while also accommodating the need to provide administrative data access for the public good.

S. Lombardi (✉)
VATT Institute for Economic Research, Helsinki, Finland
e-mail: stefano.lombardi@vatt.fi

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_10

The chapter's focus is on *administrative data*, which comprises unit-level information (typically, individuals or firms) where the units are followed over time (at the yearly, monthly or sometimes daily level). Finland, like other Nordic countries, has a relatively long history of collecting, cleaning, validating and granting access to administrative registers for survey or research purposes. For this reason, it is not feasible to try to provide a comprehensive list of databases and data providers. Instead, this chapter's aim is to explain how the system works in a very practical way and to provide details about how to get access to Finnish socioeconomic microdata in the most common and flexible way, which is via Statistics Finland, the Finnish National Statistical Authority.[1,2]

Statistics Finland's microdata can only be used for scientific research or statistical surveys and can be accessed remotely via secure two-factor authentication or in the Statistics Finland's research laboratory.[3] Statistics Finland's remote access service was launched in 2010 in order to promote easy and equal access to data from researchers' local workstations in a monitored and secure environment, thereby observing data protection rules. As outlined below, Statistics Finland provides access to a wide range of administrative registers (*ready-made* registers) and can provide *tailored* microdata based on researchers' needs. Moreover, data delivered from external organisations (including population-level microdata) and self-collected data can be linked to Statistics Finland's registers, under specific conditions. Since Statistics Finland takes care of both pseudonymising and linking the external data to the other databases used in a project, the overall system is very flexible and provides virtually endless possibilities for scientific research purposes.

The remainder of the chapter is organised in the following way. Section 2 describes different administrative data sources in Finland. Section 3 presents the application process for administrative data access at Statistics Finland. Section 4 further details the data protection rules in place. Section 5 presents some good practices for working with Finnish administrative data at Statistics Finland. Section 6 discusses equality of data access, and Sect. 7 concludes with a discussion about data access from a policy perspective.

---

[1] Statistics Finland is part of the Finnish National Statistical Service, a group of 13 organisations (mostly government agencies) that produce official statistics in Finland by making use of data that they collect for national and international use (see the website and references therein: https://stat.fi/org/tilastotoimi/index_en.html).

[2] Another large microdata provider in Finland is Findata, which has a wide selection of health administrative registers. In the interest of space, I will focus on data access in Statistics Finland's environment. However, researchers who are specifically interested in doing health-related research should also consult Findata's website on how to get a data permit. It should be borne in mind that data from Findata can be linked to Statistics Finland's data and accessed from Statistics Finland's servers, whereas the opposite is generally not the case. This also explains why the chapter focuses on accessing data via Statistics Finland.

[3] In special cases, certain data can be delivered to an organisation. Since this is not the typical and most flexible way to access data, I focus on the most likely cases in which the researcher is interested in accessing the data remotely.

2 Data Sources

All microdata registers accessed from Statistics Finland's servers can be combined by using protected unit identifiers (pseudoidentifiers) that remain static over time. This allows researchers to follow individuals over time and across registers in a privacy-preserving way.[4] Through pseudoidentifiers, workers can be linked to firms and establishments, co-workers to their peers, and individuals to spouses and other family members across generations. Firms themselves can be observed over their life cycle, and detailed information from balance sheet and personnel data can be linked over time.

This type of data can be used to study a broad range of topics in economics and social sciences. For instance, information about workers' labour market histories and training programmes can be used to analyse the effectiveness of labour market policies and the functioning of unemployment insurance systems (Kyyrä & Pesola, 2020; Kyyrä & Wilke, 2007) and of the education system (Pekkala Kerr et al., 2020). Standardised cognitive and non-cognitive test scores, linked to education and labour market outcomes across generations, can be used to analyse the role played by education reforms in shaping skills and the intergenerational transmission of income (Pekkarinen et al., 2009; Pekkala Kerr et al., 2013). Questions concerning health inequality have been studied by linking socioeconomic data to health registers, such as the birth and cause of death registers (Huttunen & Lombardi, 2021; Schwandt et al., 2021) and the cancer, hospitalisations and drug reimbursement registers (Liu et al., 2024). Following firms over time, while linking employers to employees each year, can be exploited to study questions concerning tax incentives (Harju et al., 2018; 2019; Benzarti et al., 2020), job loss (Barreto et al., 2024) and wage inequality (Palladino et al., 2025). Among the several additional registers available for research, detailed information on housing linked to individual and family characteristics can be used to analyse urban economics-related questions (Bratu et al., 2023).

Although the quality of the Finnish registers is high by international standards, it is good to keep in mind that they are not free of measurement error (duplicates, erroneous variable values and missing information) and that statistical imputation and adjustments implemented by the Statistical services, prior to their introduction of this information to the public, can also induce some measurement error. In some cases, it is not even possible to uniquely merge observations across different registers. This places much of the responsibility on researchers to look for solutions to such problems.

A useful distinction, when thinking about microdata in Statistics Finland's environment, is whether they are considered *ready-made, tailored* or *from external providers.*

Ready-made registers, as their name suggests, are registers located on Statistics Finland's servers that are ready to be used. The ready-made research data sets are quicker and cheaper to use than tailored or external data sets. This category of data

[4] No one, except for Statistics Finland, has the key to link personal identity number to the pseudoidintifier.

comprises very detailed, population-level, longitudinal information on individuals and establishments/enterprises, so that, in practice, many projects in economics and other social sciences can be performed by using this type of register data exclusively.[5] The available information (which is continuously updated and augmented in scope), includes, among other things:

- *Individual data:*
  - employment history, income and transfers sources, wage, hours worked, overtime, bonuses, occupation and industry;
  - yearly link between employer and employee via unique identifiers;
  - unemployment, programme participation and pension spell data;
  - education enrolment and achievement, test scores, matriculation exam results and qualifications.
- *Business activity data* (at the establishment and enterprise level):
  - attributes: industry, location, form and ownership;
  - activities via balance sheet information: profits, sales, import and exports, R&D expenditure;
  - personnel: wage bill, number employed, education, occupation and mobility.

The population-wide registers are typically available on a yearly basis starting from 1987 for people who are alive at the end of the year. Moreover, with the census registers one can follow people back in time up to 1950 (for a 10% sample of individuals) and the whole population between 1970 and 1985 (every five years). The data is in daily spell format in other registers, such as for employment, unemployment and programme participation.

Statistics Finland's *tailored data* consists of individual or aggregated research data produced by Statistics Finland's statisticians based on researchers' needs. Many aggregated databases are freely available to the public. Upon request, and for a payment of an hourly-based fee, it is possible to request data produced in a specific format or for certain groups of units. This information can be linked to existing ready-made registers or to other data sources obtained from external organisations.

Finally, *external data* corresponds to any data sources (including population-wide administrative registers) that are imported from outside of Statistics Finland's servers and that are linked to the existing registers at Statistics Finland. From the perspective of data costs, external data are considered tailored data, in the sense that they require the work of a statistician for their pseudonymisation and linkage via secure unique individual identifiers.

[5] Data sources available for research are constantly updated and augmented by Statistics Finland's research services. Despite this, many data sources (such as old census data, wealth information and more) still need to be digitised and carefully validated. Such demanding efforts are sometimes initiated by researchers themselves in collaboration with the data providers, so that the new data sources can be made available more broadly to other researchers thereafter.

For the purpose of linking microdata provided by external organisations, the user needs to provide Statistics Finland with a *data permit* issued by the relevant organisation, which is a legal document that states the permit to use the data and to link it to the other necessary registers for the purposes of carrying out a specific research question. Prior to performing the pseudo anonymisation, and linkage of the external data with any other data sources, Statistics Finland has the obligation to assess whether the resulting merged data sources abide by the rules in the Finnish Statistics Act (280/2004) (see Sect. 7).

Several Finnish organisations routinely provide their microdata for research purposes. Among these, for instance, there are the Ministry of Labour, the Ministry of Education, KELA (the Social Insurance Institution) and Findata. Findata, which is the data permit authority for the social and healthcare sector, deserves a special mention because it provides a unified platform through which to obtain data from different health authorities (such as KELA, the Ministry of health and more) under a single data permit. The health registers that cover the entire population are available via Findata and include the cancer register, cause of death register, hospitalisations (in-patient and out-patient data) and drug prescriptions and reimbursements, among many others. See Atella et al. (2025), who describe access to health registers in Italy, Denmark and Finland, for detailed information on access to health registers via Findata permits.

It is worth stressing that some data holders (such as the Ministry of employment and Finnish Ministry of Social Affairs and Health) have agreements with Statistics Finland so as to facilitate access to specific popular registers.[6] Therefore, even if these registers are considered external data sources, Statistics Finland can in practice provide support with the necessary data permit process prior to issuing the user license.

3 Application Process for Data Access

3.1 Overview of the Application System

It is first necessary to obtain a *user license* to access microdata on Statistics Finland's servers. The applicant needs to submit a research plan electronically with the goal of the research and the databases and sample or population of reference that are required to reach that goal. The application also needs to include a pledge of secrecy, as well as the user licenses from any external data sources that need to be linked to Statistics Finland's microdata (if any). Sensitive data, such as self-collected survey data or data generated via experiments, require the applicant to submit ethical approval from a recognised institutional review board.

[6] Information includes additional education registers, spell data on participation in labour market programmes and unemployment insurance, criminal records, date and cause of death data and more.

In general, a project at Statistics Finland can be used to answer more than one research question, but proposals that are too broad are typically not accepted, given that they likely do not comply with data minimisation requirements. A data permit lasts at most five years from the time that the permit is issued, but it is possible to extend the permit by applying for a continuation of the user license. It is important to stress that the documents signed specify clear obligations for data protection and are legally binding. Users found to be infringing the rules are subject to fines and up to one year of imprisonment. Rule infringements can also result in a lack of data access for future applications (for the user and, possibly, for their organisation). See Sect. 4 for more information on the rules governing data access and data protection.

The applications are processed on a first-come, first-served basis. The assessment of whether a user permit can be issued is based on the legislation that is currently in place (in terms of privacy rules, and in relation to the data required). In other words, despite the fact that the data permit is, ultimately, issued for research purposes, there is no scientific committee that evaluates the research proposal. People that are not satisfied with Statistics Finland's decision can amend the application or can file an appeal. If an agreement is made and signed by both parties, then the data is subsequently compiled and can be accessed via secure remote access. Statistics Finland follows the legislation on data protection and confidentiality practices when compiling and releasing the data; this implies that having a data license from an external organisation, with the purpose of working on a specific research question, does not automatically guarantee obtaining a Statistics Finland's license to open the corresponding project. Statistics Finland's licenses, the data, their tailoring and the remote access are all subject to costs (see sub-Sect. 5.4).

3.2 Conditions to Open a Project and Access Data at Statistics Finland

The application can either be submitted by an authority, institute or by individual researchers. Regardless of the applicant, each user license is always granted to a person or group of people, even in the case when the applicant is an authority or institute. It is the responsibility of the licensed user to inform the other data processors about the terms of the license and to monitor their compliance. In order to submit an application, the applicant first needs to have Haka, Suomi.fi, eduGAIN or Virtu/State Identification Service credentials.[7] The same credentials are later used to log on to remotely access the data with the FIONA two-factor authentication service via a Finnish organisation that has a remote access agreement with Statistics Finland.

[7] Haka is the identity federation for the Finnish higher education and research institutions, Suomi.fi is the identification system of residents in Finland, eduGAIN is a system providing unique identification across different countries and Virtu is the identification system of Finnish civil servants.

With respect to processing the data, only the group of people whose names are included in the license can process the data. The user license can be updated (for a fee) so as to reflect changes in the research team or in the databases required for the analyses and to extend the time validity of the permit itself.

When it comes to remotely accessing the data, the researcher needs to be affiliated (or at least have an agreement) with a Finnish institute that has a valid agreement for remote access with Statistics Finland. The restriction for the organisation or institute to be a Finnish one is motivated by data protection laws.[8] Remote connections are made through the Finnish organisation via VPN, not to Statistics Finland directly. In case an organisation does not yet have a remote access agreement with Statistics Finland, the applicant can submit their remote connection information security and data protection statements and Statistics Finland's Research Services will draw up a new agreement with the institution.

Remote data access occurs by opening a remote network connection from the researcher's local workstation to Statistics Finland's server. The remote connection through the FIONA system can only occur from the IP addresses specified by the customer for the users' workstations. Researchers working from outside of Finland can use remote access through a Finnish organisation with a valid remote access agreement. Access from abroad requires an approved user license decision; this can be done when opening the project or later via a user license update. Remote access use is possible only in EU and EEA countries and in those non-EU countries in which the level of data protection has been recognised as sufficient by the EU Commission. Researchers commit to ensuring that computers, networks and work premises are data secure. Working from home is possible as long as work can take place with a VPN connection through the employer's network. The mobile phone, through which the two-factor authentication takes place, must also be stored carefully.

In conclusion, the precondition to be able to work remotely with administrative data at Statistics Finland is to have an affiliation or an agreement with a Finnish institution. Having an affiliation with a Finnish research institute recognised by Statistics Finland provides full flexibility to both open a new project and access the data remotely via the FIONA secure system. Researchers based in non-Finnish research institutes need a research agreement with a recognised Finnish institution as it is not common for Finnish research institutes to provide external affiliations to foreign-based researchers. In practice, such agreements are drawn up when researchers based in non-Finnish research institutes start a collaboration with Finnish-based researchers.

[8] The processing of personal data is guided by the General Data Protection Regulation (2016/679) of the European Union and by the supplementary national Data Protection Act (1050/2018).

3.3 Scope of the User License and of the Administrative Registers

The assessment of how much data is possible to apply for with a given user license is always made on a case-by-case basis. In principle, the data source (ready-made, tailored or external) is not central for the final decision of issuing a user license. What matters are data protection considerations about whether the final set of data, when linked together, is in line with the current legislation (see subsequent section). This means that, in principle, it is possible to flexibly link together Statistics Finland's registers together with external administrative data, provided that the data linkage is justified by the research question and is in line with the data protection legislation.

As mentioned previously, several external registers can be relatively easily linked with ready-made registers, and Statistics Finland even facilitates the external data request and handles the data permits required. Other registers can still be linked but require a more expensive and time-consuming external application process (such as for health data via Findata). While the system is extremely flexible by international standards, it is good to keep in mind two reasons that can limit the linkage of external data sources. First, understandably, the institution that has the right to provide the permit to use the external data always retains a certain degree of discretion in providing the data permit. Second, Statistics Finland's obligation to analyse whether a user license is in line with data protection rules in practice translates into more in-depth assessments whenever an application requires external data to be linked to ready-made data sources.

4 Data Protection

Population-wide administrative registers can be accessed in a relatively straightforward way in Nordic countries thanks to a well-developed body of laws and trust in public institutions. This section's rationale is to provide some background information on how scientific research that uses microdata, which is a task which is in the public interest, can be performed in the Finnish context while limiting the disclosure of sensitive information. While a comprehensive review and assessment of data protection goes beyond the scope of this chapter, this section presents some of the Finnish system's key data protection elements. Both Statistics Finland and researchers are responsible for guaranteeing data protection. The focus of this section is on those elements with which researchers must comply.

4.1 User License Terms and Conditions

Data protection laws at both the National and the European levels govern the extent to which microdata information can be accessed and used, all the way from the beginning of the application process up to the ways in which results can be exported from Statistics Finland's servers and thereafter disclosed to the public.

When signing a user license agreement, the researchers must comply to use the data solely for the purposes specified in the user license, which must be scientific research or must concern statistical surveys about social conditions. The data must be treated as confidential and cannot be disclosed for any reasons at each step in the workflow. Users cannot attempt to identify data subjects in the data and must comply with the pledge of secrecy that they sign prior to accessing the data.

The use of the data is conditional on having a valid license throughout the processing period; the same holds for data permits required to use any external data sources. When the data license expires, the researchers lose access to the remote access system, and the obligation of secrecy remains in effect even after the user license expires. Moreover, the user may be asked to provide a presentation of their results before the research is published and copies of all research output must be sent to Statistics Finland.

4.2 Remote Access and Exporting Output Data

The primary and most flexible way by which to access a project at Statistics Finland is remotely via the FIONA secure two-factor authentication system. The rules governing data processing and data transfer outside Statistics Finland's servers aim to protect the material published so that no individual persons or enterprises can be identified. Research results and other materials can be exported from the Statistics Finland's servers via a drop folder system, and the output being exported is subject to random checks.

Statistics Finland provides detailed rules, examples and support materials on how tabular data, data in graphical format or model estimates should be exported. Two common sensitivity rules are as follows:

- *Threshold value*: a cell is considered sensitive if it includes fewer statistical units than a predetermined threshold value. The threshold is generally 3;
- *Dominance rule* (or *n, k rule*): a cell is considered sensitive if the *n* largest statistical units amount to at least *k* per cent of the cell's total value.[9]

It is clear from the above that minimum and maximum values of continuously measured variables (such as income), for instance, cannot be exported, whereas

[9] Statistics Finland's research services specify that business data with fewer than 15 months from the reference date should use (n, k) values of (1, 75). See references at the end of the chapter.

proportions can usually be exported as long as they are computed by using at least three units. In graphs, the presence of outliers must always be masked, and so on.

5 Good Practices for Working with Finnish Administrative Registers

5.1 Read the Metadata Before Starting the Application Process

It is not always obvious which database should be used for a given purpose, given the large number of databases available at Statistics Finland and because certain types of analyses require complex data management of several population-wide administrative registers. The data catalogue on Statistics Finland's website provides the documentation of all databases available. The most frequently used databases have documentation in English, while the others have documentation files in Finnish exclusively. While the pdf documentation files can be software-translated easily, the result is not always perfect.

Reading the documentation and collaborating with experienced users can prove valuable prior to the initiation of the application process. Moreover, the staff at Statistics Finland are always available for data-related requests. Statistics Finland has data experts specialised in the use and maintenance of different registers for the most in-depth and complicated questions, so that it is often possible to obtain satisfactory answers.

5.2 Consider If Joining an Existing Project Is a Viable Possibility

Depending on the type of research question, it might be possible to join an existing project without having to open a new one. In fact, a user license is always limited in the scope of its research plan, but it can, in principle, encompass related research questions, which in turn makes it possible to join already-opened projects. If in the scope of the research plan, additional registers can be added alongside the addition of the new researcher's name on the license.

5.3 Understand the Queuing System

Statistics Finland has separate queue systems. Small changes to existing user licenses (such as adding new researchers to the license or extending the access period) and

applications for user licenses that require access to ready-made databases are put on fast queues, usually requiring about one month to be processed. Conversely, applications that include tailored data or the pseudo anonymisation and linkage of external data sources require a longer wait time to be processed (approximately 4–6 months, and sometimes longer).

Given the above, it is possible to open the project by requesting access to ready-made registers only (therefore, being put in the fast queue) and adding the external data to the license by updating it later, provided that the analysis requires external data but not at the immediate start of the project. This allows users to save time and to start preparing the data and to perform some of the analyses. In some cases, this approach is even necessary because the institution providing the external data has not yet provided the data permit for the external data, which might take a long time to be issued.[10]

5.4 Project-Related Costs

Opening and maintaining a project at Statistics Finland can be expensive and costs have been rising in recent years. The researchers undersigning the user license agree to pay for different types of costs[11]:

- *One-off costs*:
  - User license (unless the license needs to be updated);
  - Ready-made data subject to a one-off fee per register applied for.
- *Yearly costs:*
  - Annual fee for continuously delivered ready-made data modules;
  - FIONA annual fee and remote access fee per user-year; FIONA virtual machine capacity charge.
  - Additional disc space and software.

While there is little to do about the cost for remote access, it is easy to limit the amount of money spent by carefully choosing which databases to request. Moreover, in many cases, having continuous updates is not always strictly necessary, and either way one possibility is to require data that is subject to a one-off fee and then, at a later stage, to apply for continuously updated registers where necessary (by paying

[10] Despite this, it is good to keep in mind that an already issued data permit cannot be bended to include new, loosely related research questions. In such cases the user needs to open or join a brand-new project.

[11] As of January 2026, one-off costs are: user license (800–3,200 EUR, with additional 115 EUR/hour for licenses requiring more than 15 h of work); ready-made register fees (800–900 EUR per database). Yearly costs: annual fee for any continuously delivered ready-made data module (600–700 EUR per year-database); FIONA annual fee (3000 EUR), FIONA remote access (460–3680 EUR, with an additional 500 EUR per user-year); additional disk space (200 Gb): 120 EUR/year. With the exception of the user license cost, an additional 25.5% VAT must be paid on all these costs.

a fee to re-issue the permit). The converse is also possible: to start with continuously updated registers and later to switch to one-off registers (retaining the continuously updated data paid for up to that point).

Overall, a user license allows researchers to apply for a combination of continuously updated and one-off databases, which is often a sensible starting point, depending on the analysis that is to be performed. It is also good to keep in mind that it is always possible to modify a permit in order to require additional databases (and to add more researchers to the permit) at a later stage. However, modifying a permit many times is not practical, as it is costly and time-consuming. Therefore, researchers should always have a structured data minimisation and research plan in mind upon starting the application process.

6 Equality of Data Access

Overall, the system appears to be both well developed and equal. It certainly helps that the system is centred around the national statistical authority and has clear rules governing access to the data remotely in a secure environment. Moreover, the decision about whether a given permit can be delivered is solely based on data protection rules and the application is evaluated only on the basis of its legal feasibility. The same holds true for the decision about whether microdata external to Statistics Finland's servers can be linked to the already available registers. The resulting combination of registers that can be used for research purposes is virtually endless, conditional on data protection rules being in place. Finally, the data can easily and flexibly be accessed from any EU and EEA country (and from additional countries which are recognised as secure, according to EU regulations).

From the perspective of non-Finland-based researchers, how easy is it to access administrative registers? Due to the implementation of the GDPR, remote access can only occur through secure connection via a Finnish institution that is recognised by Statistics Finland. In practice, this implies that a researcher who is interested in using Finnish administrative registers needs to be part of a research team that includes someone with a Finnish institute affiliation. While at first glance this might appear restrictive to foreign-based researchers, in practice it is sufficient to start a collaboration with a researcher who is affiliated with a Finnish institution. Keeping the remote access agreements between Statistics Finland and research centres a national matter makes Statistics Finland's gatekeeper role practical and feasible. Moreover, given the level of complexity of the administrative registers and the knowledge required about administrative details, collaborating with people affiliated with Finnish institutions to work on Finnish population-wide registers would be very strongly advised, even if it were not required by law. Therefore, this appears to be a reasonable rule that

other countries, interested in opening access to administrative registers, could draw inspiration from.[12]

Researchers need to have a data permit from the data holder when it comes to linking external data to Statistics Finland's ready-made registers. In many cases, getting the necessary data permit can be surprisingly easy, provided that the research question is in line with the interests of the data holder. Other data sources might be harder to obtain without having first built trust with the given institution. This is understandable given that the data holder providing the data permit is, by law, a gatekeeper who is intended to monitor against data misuse. In this case, it also might prove useful (and often necessary) to work with researchers that have an affiliation with a Finnish institution. Again, this goes hand in hand with the necessity of navigating the crucial data-related and institutional details required to use the data in a proper manner.

Knowledge of the institutions is more broadly necessary in order to correctly clean and manage all Finnish administrative registers. While the quality of Statistics Finland's documentation is generally high, and ever-improving, metadata documentation does not always exist in English. This might discourage certain researchers from trying to obtain access to Finnish administrative registers. However, questions and lacks in the documentation can always be brought up with the research services at Statistics Finland, who have a strong incentive (motivated by the law) to provide clear answers to both current and potential data users.

One factor that can increase the risk of unequal data access is data access costs. In recent years, opening and maintaining a project has become quite expensive, especially when continuously updated registers are required and/or external data sources are linked to the ready-made registers. With the goal of increasing revenues to cover the costs of producing research services, substantial price increases occurred in 2022 (an average increase of 47%) and 2026 (ranging from 10% to 30%), with a further 45% average increase expected in 2027. Since grant money is limited and is typically allocated based on previous performance, if this results in more senior and established researchers being able to get funds, this type of gatekeeping might reduce junior researchers' ability to access the data, unless they have connections with more senior researchers in Finland. While it is possible to apply for grants covering data or other costs (both from public and private organisations and foundations) in Finland, spikes in data costs nonetheless risk reducing resources available to younger, less well-established researchers, and to researchers more broadly.

[12] The other Nordic Countries use a similar approach when it comes to implementing the GDPR (see, for instance, the chapter on Sweden).

7 Conclusions and Way Ahead

This chapter has provided an overview of how to access administrative data in Finland, with a specific focus on the most common way to do it, from the perspective of social scientists at least, which is remotely and via a user license at Statistics Finland. Finland, alongside with other Nordic countries, is at the forefront of providing administrative data for research purposes and can provide a useful benchmark for other European countries (see, e.g., Sovala, 2025). I will conclude with a discussion of the following two policy-relevant questions: *can the Finnish case provide an example to other countries interested in opening access to administrative data? What can policymakers do to make the access to administrative registers for research purposes possible?*

This discussion must start from the consideration of data protection rules. In fact, data access can only occur when appropriate data protection rules are in place. After all, this book provides first-hand evidence of how countries sometimes interpret the GDPR very differently, and often justify the impossibility of linking or providing access to administrative data due to the GDPR. Therefore, the first step in discussing how to provide access to administrative data would be to involve lawyers specialised in data protection and in the implementation of the GDPR into the discussion. In practice, a delicate balance must be stricken between opposing interests and rights. Data must be protected from misuse and privacy must be guaranteed. At the same time, administrative data access can lead to quality and policy-relevant research, which is a public good.

The Finnish case shows that the balance between tasks of public interest and data protection can be achieved in agreement with National and EU laws, while allowing researchers to access very detailed, individual-level information. Other EU member states must abide, as Finland does, by the GDPR and can even take inspiration from the Finnish system when considering whether to open administrative data access for research purposes. At the same time, policymakers should also be aware that importing the Nordic model into other settings is only likely to work in a stepwise fashion, because the institutional context also matters. The current status quo was reached not just by developing advanced data protection laws that are in line with EU-level regulations: a great deal of time and effort has been put into building trust in the institutions and into pushing the idea that it is in the public interest to evaluate public policies and to do scientific research by using high-quality administrative registers.

What can Finnish policymakers do to improve access to administrative socioeconomic registers? While this chapter has described a system that can be considered as ideal in many respects, Finnish policymakers should consider the fact that Statistics Finland holds a de-facto monopoly position when it comes to access to socioeconomic registers and to pseudonymise and link different microdata sources. Since researchers who are interested in using Finnish socioeconomic registers do not have alternatives to access the data, it is important to make sure that the increasing data costs do not undermine equality of access to the data. In particular, because research

based on administrative data generates substantial public benefits beyond those accruing to individual researchers or research institutions, the pricing of the research services should strike a better balance between the goal of full cost recovery and the wider social value of broad and equitable access to administrative data.

8 Contacts, Useful Links and Relevant Legislation

Microdata page and research catalogue:
https://stat.fi/en/services/services-for-researchers
https://taika.stat.fi/en/index
Statistics Finland's pricing of research data and services:
https://stat.fi/en/services/services-for-researchers/price-list-and-invoicing
Ministry of Finance's regulation on the payment of Statistics Finland services:
https://www.finlex.fi/fi/laki/alkup/2023/20231220
Rules of Statistics Finland Research services and instructions for licensing service:
https://www.stat.fi/media/uploads/tup_en/mikroaineistot/instructions_for_licensing_service.pdf
Data protection and output checking process in remote access use:
https://www.stat.fi/media/uploads/tup_en/mikroaineistot/fionan_tietosuojaohjeet_en.pdf
https://guides.stat.fi/remote-access-to-research-data/statistical-data-protection
https://stat.fi/meta/tietosuoja/index_en.html
Statistics Act (280/2004) and links to EU statistical legal acts:
https://www.stat.fi/meta/lait/index_en.html
https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en
Statistics Finland's contact for questions concerning research data: tutkijapalvelut@stat.fi
Findata data permits: https://findata.fi/en/permits/data-permits/

Acknowledgements I would like to thank Cristina Bratu (VATT Institute for Economic Research) and Marianne Johnson (Statistics Finland) for useful comments and discussions.

References

Atella, V., Ganna, A., & Lombardi, S. (2025). Optimizing population health through strategic use of health data. *Economia Italiana.*

Barreto, C., Fluchtmann, J., Hijzen, A., Lombardi, S., Bennett, P., Bertheau, A., Chan, W., Gorshkov, A., Hambur, J., Johnstone, N., Lochner, B., Meekes, J., Mehdi, T., Muraközy, B., Nolan, G., Salvanes, K., Nordström Skans, O., & Vejlin, R. (2024). *"The "clean energy transition" and the cost of job displacement in energy-intensive industries.* OECD Social, Employment and Migration Working Papers, No. 310, OECD Publishing, Paris

Benzarti, Y., Carloni, D., Harju, J., & Kosonen, T. (2020). What goes up may not come down: Asymmetric incidence of value-added taxes. *Journal of Political Economy, 128*(12), 4438–4474.

Bratu, C., Harjunen, O., & Saarimaa, T. (2023). JUE Insight: City-wide effects of new housing supply: Evidence from moving chains. *Journal of Urban Economics, 133*, Article 103528.

Harju, J., Kosonen, T., & Skans, O. N. (2018). Firm types, price-setting strategies, and consumption-tax incidence. *Journal of Public Economics, 165*, 48–72.

Harju, J., Matikka, T., & Rauhanen, T. (2019). Compliance costs vs. tax incentives: Why do entrepreneurs respond to size-based regulations? *Journal of Public Economics, 173*, 139–164.

Huttunen, K., & Lombardi, S. (2021). Mortality inequality in Finland. *Fiscal Studies, 42*(1), 223–244.

Kerr, S. P., Pekkarinen, T., Sarvimäki, M., & Uusitalo, R. (2020). Post-secondary education and information on labor market prospects: A randomized field experiment. *Labour Economics, 66*, Article 101888.

Kyyrä, T., & Wilke, R. A. (2007). Reduction in the long-term unemployment of the elderly: A success story from Finland. *Journal of the European Economic Association, 5*(1), 154–182.

Kyyrä, T., & Pesola, H. (2020). Long-term effects of extended unemployment benefits for older workers. *Labour Economics, 62*, Article 101777.

Liu, A., Akimova, E. T., Ding, X., Jukarainen, S., Vartiainen, P., Kiiskinen, T., Koskelainen, S., Havulinna, A. S., Gissler, M., Lombardi, S., Fall, T., Mills, M. C., & Ganna, A. (2024). Evidence from Finland and Sweden on the relationship between early-life diseases and lifetime childlessness in men and women. *Nature Human Behaviour, 8*(2), 276–287.

Palladino, M. G., Bertheau, A., Hijzen, A., Kunze, A., Barreto, C., Gülümser, D., Lachowska, M., Lassen, A. S., Lombardi, S., Meekes, J., Murakozy, B., & Skans, O. N. (2025). *The impact of firms on the relative pay of women across countries.* VATT Working Papers 181.

Pekkala Kerr, S., Pekkarinen, T., & Uusitalo, R. (2013). School tracking and development of cognitive skills. *Journal of Labor Economics, 31*(3), 577–602.

Pekkarinen, T., Uusitalo, R., & Kerr, S. (2009). School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform. *Journal of Public Economics, 93*(7–8), 965–973.

Schwandt, H., Currie, J., Bär, M., Banks, J., Bertoli, P., Bütikofer, A., Cattan, S., Chao, B. Z. Y., Costa, C., González, L., Grembi, V., & Wuppermann, A. (2021). Inequality in mortality between Black and White Americans by age, place, and cause and in comparison to Europe, 1990 to 2018. *Proceedings of the National Academy of Sciences*, *118*(40), e2104684118.

Sovala, M. (2025). Rekisterit mahdollistavat tietopohjaisen yhteiskuntapolitiikan. In J. Saari (Ed.), *Hyviä uutisia Suomesta -Menestyvän yhteiskunnan tilannekuva* (pp. 249–264). Vastapaino.

Stefano Lombardi is a Senior Researcher at the VATT Institute for Economic Research (Helsinki). He received his Ph.D. in Economics at Uppsala University in 2019, while being affiliated with IFAU (Institute for Labour Market and Education Policy Evaluation). Stefano had previously earned a master's degree in Economics from Bocconi University. He is affiliated with IFAU, IZA (Institute of Labour Economics) and UCLS (Uppsala Center for Labour Studies). Stefano's research interests are in applied labour and health economics, with a particular focus on the evaluation of labour market policies, the study of the consequences of job loss and the analysis of wage inequality. For his research, Stefano uses population-wide administrative registers from both Sweden and Finland. His research has been published in prestigious journals, such as in the *American Economic Review: Insights*, *Nature Human Behaviour* and *Proceedings of the National Academy of Sciences.*

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 11
Access to French Administrative Micro-Data



Brice Fabre and Kamel Gadouche

Abstract This chapter documents access to French administrative microdata for the purposes of research. We present an overview of both the rules and formal procedures, the concrete implementation of this access, and perspectives for potential improvement. The institutional framework for accessing microdata in France has evolved as an attempt to reach an equilibrium between the decision-making power of data producers, informal control, and the need to develop a reliable public infrastructure by which to access sensitive data. We point out this system's successes to date, which has strongly increased access to precise and relevant data for the purposes of research. Still, room for improvement remains, especially concerning data matching and results replication.

1 Introduction

Because administrative micro-data sources give private sensitive information, gaining access in France is highly protected by law, through "statistical secrecy," which defines thresholds of aggregation that need to be reached for publication for each kind of data.[1] By default, information that does not reach these thresholds is not available outside the confines of the data producer.

[1] For fiscal data, for instance, an aggregated number has to be produced with at least 11 observations, with none of them representing more than 85% of the aggregate, in order for the data to be made publicly available.

In this chapter, the denomination "administrative micro-data" covers both raw micro-data, used by the administrations for their prerogatives, and micro-data that is produced by public institutes of statistics from administrative sources.

B. Fabre (✉)
Institut des politiques publiques (IPP), Paris, France
e-mail: fabrebri@gmail.com

K. Gadouche
Secure Data Access Centre (CASD), Paris, France
e-mail: kamel.gadouche@casd.eu

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_11

While this legal feature had prevented researchers from gaining access to individual administrative data, a law in 2008 introduced an exception for the purposes of research.[2] The underlying idea was that researchers were not interested in one or even several individuals per se, but were instead interested in the global picture that administrative datasets can provide in terms of knowledge for the advancement of science and society.

This chapter aims to document this system and the resulting practice. It is structured as follows: Sect. 2 presents the rules and procedures governing access to French administrative micro-data. Section 3 describes the main public infrastructure developed in France that exists to provide access to these data. Section 4 presents current rules in terms of matching administrative databases and opportunities for improvement. Section 5 examines the difficulties in reconciling data secrecy needs and replication for peer-reviewing. Section 6 concludes.

2 Rules and Procedures

This section describes the rules and procedures governing access to *individual* administrative data (information at the individual level, the household level, the firm level, etc.) for research purposes in France. These rules and procedures are characterised by: (i) a collegial monitoring of data access, conducted by the "Committee of statistical secrecy," with necessary approval from the data producer (the administration) and (ii) the development of a shared infrastructure for access to sensitive data. This model has leaded to **a gradual and important rise in administrative micro-data uses in research overall**, especially over the course of the last decade.

More precisely, researchers can gain access to administrative micro-data under specific procedures, involving different actors:

- The research team (and its organisation of affiliation) requesting access. Researchers have to be either EU citizens or members of a project involving at least one European institution or must be located in the EU.
- The administrations that make the decisions related to the request: (i) the French administration producing the requested data sources (the "data producers") and (ii) the public service of national archives (*Service interministériel des Archives de France*, SIAF).[3]
- The "Committee of statistical secrecy" (*Comité du secret statistique*, CSS), consulted about each request. It provides an opinion and it is allowed to request a formal justification of their decision from data producers.

[2] Law No. 2008–696 of July 15, 2008: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000019198529.

[3] The competency of managing public archives is allocated to the SIAF for the main part of public data. This competency is allocated to other services for some data sources; for the purpose of simplification, "SIAF" will refer to all services having this competency for any source of data in this chapter.

- The organisation implementing the access, in cases where the request is accepted, that is all security conditions implied by the necessities to observe statistical secrecy. For most accesses, this mission is fulfilled by the "Centre for secured data access" (*Centre d'accès sécurisé aux données*, CASD), a public organisation specifically devoted to this task (data producers can decide to provide the secured access themselves, but it is often an exception).

Accesses are **project-oriented**, not field-oriented (it is not possible for a research department to have permanent access to a given data source because it is relevant for research in its field). The idea is to provide a secured environment to researchers with all data sources claimed for a given project. The access ends when the project ends.

2.1 The Different Steps to Ask for Access Permission

Request procedures are carried out by the CSS. The procedure can be decomposed between the following steps[4]:

- **Project submission:** the research team has to fill an online form, from the "Confidential Data Access Portal" (CDAP), that describes the research project, lists the data sources claimed, and outlines the reasons why these sources are necessary for the project.[5]
- **Data producers' validation for examination:** the form is sent to data producers of all the sources claimed. The application has to be validated by all of the data producers in order to be examined by the CSS. What data producers validate at this stage is their agreement that the CSS examines the application, not the agreement for the project.
- **The CSS examination:** the CSS has to provide an opinion, and data producers (which also have their seat in the CSS) make the decision for those data sources for which they are responsible. The CSS is composed of data producers, representatives of researchers and representatives of many other actors: the Parliament, the SIAF, employees' and employers' unions, the central administration and local governments. The presidency of the CSS goes to a judge of the *Conseil d'Etat*, the highest authority of French administrative justice. In cases where the access request is refused by a producer, the CSS can require to data producers a precise justification of their decision.
- **The final decision:** the final decision goes to the SIAF, which receives the opinion of the CSS, in cases where the request gets a favourable opinion from the CSS and the agreement from data producers.

[4] This procedure is not substitutable to GDPR requirements. The research team is responsible for carrying out all of the procedures related to this regulation.

[5] The online platform to claim data access can be found here: https://cdap.casd.eu/comite-secret-statistique.

In case of a refusal from this procedure, the research team can contest the decision with the "Commission for access to administrative documents" (*Commission d'accès aux documents administratifs*, CADA), a public, independent administration whose mission is to ensure access to public documents under the existing legal provisions.

Information related to this general procedure is available online, in French and in English, and all research teams are subject to the same CSS meetings schedule; the request is allocated to the closest available slot after the online application is filled out by the team. There is usually one meeting every three months for new research projects examinations (and the deadline for project submission for a given meeting is one month before the meeting).[6] The SIAF decision then takes about three weeks; research teams face these same conditions in terms of permissions request. Overall, the time span between the project submission and effective access varies from between two and five months depending on the scheduled date of the subsequent CSS meeting.

This whole procedure is not specific to research, but also covers public service more generally. For instance, a public administration seeking to gain access to a data source produced by another administration is also submitted to this procedure. What is specific to research is the scientific project-orientation feature. For instance, a statistical service of the Ministry of the Economy would need access to labour contracts data (produced by the Ministry of employment) in order to provide information about the flow diagnostics related to the labour market to the minister's cabinet. In accordance with this task, this statistical service can ask for an access to these data, without there being a link to a specific project.

2.2 The Implementation of Data Access

A data access request that has received all of the necessary permissions has to be implemented on the basis of conditions aligned with the requirements of statistical secrecy. Data cannot be viewed by anyone else other than entitled researchers, and every output that researchers would like to publish requires that the statistical secrecy conditions of aggregation are met. The French national statistical institute, called Insee (*Institut national de la statistique et des études économiques*), has developed CASD in order to offer a common infrastructure by which to access any administrative source with all of the necessary conditions in terms of security. Data producers can still decide to provide this environment themselves. However, the CASD is now close to being considered the rule, rather than the exception. The following section describes this service and the daily features for data access for researchers.

[6] Researchers' requests to amend existing projects (asking for a new data source or to include a new researcher in the project) also require a CSS examination and are subject to the same general procedure. There is a CSS meeting every one- or two-months for these "complementary requests."

3 Data Access Through CASD

CASD is a public organisation responsible for implementing the practical procedures by which to access data and protects confidentiality. Its goal is to host and to make available sensitive data in a secured infrastructure. Apart from the substantial economies of scale that this shared infrastructure implies, it also allows researchers to work simultaneously on confidential data sources from different data producers and, therefore, in different sectors (which would obviously not be possible if each administration requested to use their data on their own server). At the time of writing, **CASD hosts 598 databases, representing several hundred terabytes, from diverse administrations**: Insee (the French national institute of statistics), the Ministries of Finance, Justice, Labour, Agriculture, Health, Ecological Transition, etc. The exhaustive list of sources hosted by CASD can be consulted online: https://www.casd.eu/en/data-used-at-casd/. This webpage gives a short description in French and in English for each data source, and usually a list of variables (only in French, although a translation is in development). The full documentation of databases is usually available to researchers in their secured CASD environment (in French, not in English at present). **CASD hosts around 800 scientific projects**, carried out by academic researchers, data scientists from the public sector, or from the private sector (but still with a public service mission). Figure 11.1 represents the evolution of the CASD's activity between 2012 and 2025, measured by the number of active users and the number of data sources available in the CASD, and illustrates the recent growth of administrative micro-data in research inputs. Figure 11.2 represents the geographical distribution of CASD uses across countries.

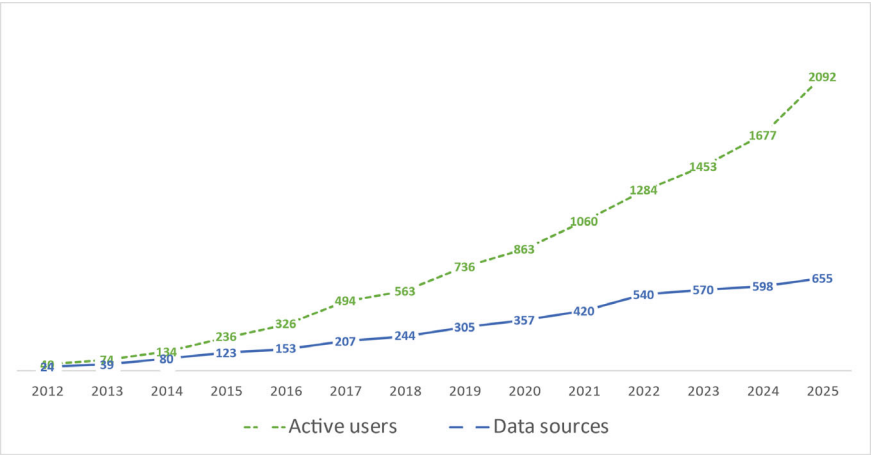


Fig. 11.1 The evolution of the number of CASD active users and data sources available in the CASD

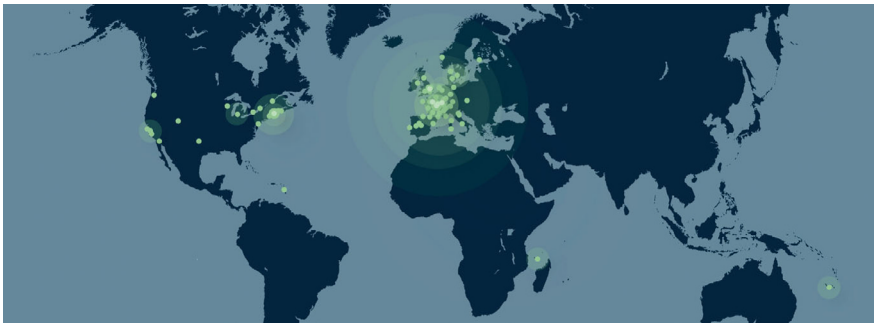


Fig. 11.2 SD-Boxes deployment map

3.1 The CASD Infrastructure

Researchers granted permission for access through CASD for a specific project are allocated an environment for this project by CASD, with all of the requested data therein. This environment is a bubble, one isolated from the internet. The researcher has access to it from his or her office through a dedicated access point: the SD-Box is equipped with both a smart card reader and a fingerprint reader. Researchers are given a personal card and have to register their fingerprint. The SD-Box simply has to be connected to the wired network through Ethernet.

This SD-box system and all of the related infrastructure, developed and patented by CASD, constitute a secure access technology, one specifically designed for end-to-end security, and provides a turnkey security solution, thereby freeing users from the need to have their institution's IT security certified. CASD's services are certified for ISO 27001 information security, ISO 27701 personal data protection (GDPR) and health data hosting standards. CASD's IT infrastructure is hosted in France, and is regularly audited by ethical hackers from ANSSI[7]-qualified companies. CASD was the first data hosting service in Europe to obtain the official GDPR certification (Art. 42) through Europrivacy in May 2025.

Gaining access permission and effectively benefitting from it are treated separately, in the sense that once permissions are obtained, the research team has to formalise its access with CASD, in a separate procedure and through an annual contract. Effective access is subject to user fees, depending on the technical configuration requested (RAM, processor, etc.), the number of users and the number of SD-boxes required. All tariffs are described online.[8] CASD's budget is based on these fees (85%) and on public funding (15%). Gaining access permission is not related to any difference of opportunity between research labs; the capacity of researchers to benefit from this infrastructure depends only on their financial capacity to pay for

[7] ANSSI (*Agence Nationale de la Sécurité des Systèmes d'Information*): the French agency for information systems security.

[8] https://www.casd.eu/tarifs-2/

the service. For instance, in January 2025, access for one researcher with the basic technical configuration (8 GB RAM, 200 GB of storage and 2 vCPU processor) called for an annual fee of 3,480 euros, while access for five users with a stronger configuration (32 GB RAM, 1 TB of storage and a 6 vCPU processor) represented yearly price of 10,800 euros.

3.2 Output Checking

In addition to providing secure access, CASD is also in charge of moving the results researchers want to publish out of the bubble. For each set of statistics that a researcher wants to move out of the secured environment, CASD has to ensure that these results fulfil the statistical secrecy conditions in terms of aggregation so as to guarantee that none of these results could allow for the linkage of a given number to an identified individual. Researchers have to provide their statistics (means, quantiles, results from regressions, etc.) with the production code, a short documentation and the number of observations used for each statistic for this precise purpose. The enrolment sessions users must complete before accessing the data include training on these rules of aggregation.

Concretely, for each researcher requirement, the CASD staff controls manually whether the numbers respect statistical secrecy conditions to be published, and move them out of the bubble provided that these conditions are fulfilled. Ultimately, the responsibility for respecting statistical secrecy goes to the researcher and the research institution.

Manual control is the most common option. Still, some projects hosted by CASD can ask for "automatic outputs" where manual checking is not systematic. This requires the permission of every data producer involved; in this case, deciding whether to check a given output manually or not is based on a machine learning model developed by CASD, which allocates a "risk score" to each output that is based on a continuously increasing database on previous outputs (the decision for this output, confronted with different output characteristics).

CASD also has a sanction system that ensures legal and personal accountability for users. After investigation, all breaches are subject to actions that are undertaken in close collaboration with the data producers and/or competent authorities.

3.3 Session Recording and Traceability

CASD has developed a programme that keeps videos of computer sessions (everything visible to the user is preserved) in order to ensure traceability, with a frequency that varies depending on the project. The same applies to IT administrator sessions. These videos are archived and sealed by blockchain on a daily basis. In exceptional cases, CASD may rely on recordings, typically during an inspection by a data

protection authority (CNIL in France). Recently, CASD developed a programme, as a proof of concept, to monitor the contents of the videos, using computer vision and text recognition technologies.

4 Matching of Administrative Data

The very wide range of administrative micro-data available at CASD opens the door to many opportunities in terms of matching administrative data from different sectors, and in terms of research potential (e.g. matching education data and employment data to investigate the links between an individual's educational trajectory and subsequent career path; matching household tax data, firm ownership and firm accounting data to measure undistributed capital income in households' economic income).

A law, dating back 2016, aimed at simplifying the procedure to match administrative micro-datasets, while also preserving the same level of personal data protection.[9] Technically, the main part of French administrative data relating to individuals could be matched through the "NIR" (*Numéro d'inscription au repertoire*), also called the "social security number" (*numéro de sécurité sociale*). This ID has to be reported by individuals in almost every administrative form. However, this ID is strongly meaningful, given that it provides information about the individual's sex and both the date and their place of birth; this is why access to the NIR is highly restrictive. Even administrations that have access thereto, for their own data, cannot have it for data beyond their administrative competency, so that they cannot make any precise match between data from different producers.

The idea of this 2016 law was to define procedures in order to match data without directly using the NIR. More precisely, there are two distinct matching systems: one for public statistics and the other for public scientific research.

1. For public statistics, the aim is to build a non-meaningful matching key from the NIR. This matching key, called the "non-significant statistical code" (*Code statistique non signifiant*, CSNS), is common to all statistical sources. This key must be changed every 10 years for the whole stock of micro-data. The choice of a non-meaningful statistical code is intended to reinforce the principle of compartmentalisation between administrative use and official statistical purposes, thereby limiting the use of data containing this code to the official statistical service alone.
2. For public scientific research, the aim is to enable data matching based on a non-meaningful matching key which is, contrary to the CSNS, research project-specific. A distinct matching key must be produced for each data-matching project. The matching result must only be accessed from a secure environment (such as CASD). This project-specific key can be produced by a "trusted third party" (distinct from the research team and data producers) and it also has to be renewed every 10 years.

[9] Law No. 2016–1321 of October 7, 2016: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746.

To sum up, current rules allow for many opportunities in terms of data matching. Still, a given matching requires the permission of data producers, and matching takes additional staff resources from administrations, which can be a reason for refusal. Finally, beyond rules, there is still room for improvement from a technical point of view, such as through the development of a robust methodology of reference to implement project-specific keys for research. In addition to improving reliability, such methodological improvements would also bring about economies of scale and would decrease the resource constraints that administrations currently face when undertaking these matchings.

5 Data Secrecy and Replication for Peer-Reviewing

The reproducibility of research on confidential data is a major issue for researchers who submit their work for publication in peer-reviewed journals. Highly restricted data do not, in practice, enable a referee to check whether the results presented in an article are reproducible, because of their legal status and the access procedures involved. In 2019, the CASD and the cascad research certification agency (UMS 2007, CNRS, HEC Paris, Université d'Orléans) initiated a pilot to carry out reproducibility checks on research results generated on the basis of data hosted on the CASD upstream of publication and providing, where appropriate, a reproducibility certificate to the applicant, who can then pass it on when submitting their work for publication. The results of this pilot were published in *Science* in July 2019 (Pérignon et al., 2019).

This pilot turned out to validate the procedure after five certifications. However, providing this service on a more systematic basis, and at a reduced cost for researchers, requires additional investments in order to automate verification procedures and to implement a dedicated infrastructure to carry out certifications, with dedicated servers, storage of environments, sealing, and archiving.

6 Conclusion

Access to French administrative micro-data for research involves different actors and has the objective of reaching an equilibrium between the decision-making power of data producers, the informal control of a larger set of representatives and a proactive policy for the development of a shared and efficient infrastructure for the provision of access to sensitive data. This institutional structure, initiated in 2008, has leaded to an important rise in administrative data uses in research, as illustrated by the quantity of data and the number of projects hosted by CASD.

Still, room for improvement remains. Although important investments have recently been made, French administrative data mainly takes the form of different and

separate sources; therefore, making matching less burdensome would be an important step in this process. Moreover, while the main part of administration is entered in the CSS process, some administrative organisations still manage access to their data through their own procedures. It is of the utmost importance to convince these institutions to join the dynamics. This system, mainly based on cooperation, is a gradual process that will take time, with the end of creating a dialogue for a more sustainable equilibrium.

Reference

Pérignon, C., Gadouche, K., Hurlin, C., Silberman, R., & Debonnel, E. (2019). Certify reproducibility with confidential data. *Science, 365*(6449), 127–128.

Brice Fabre is a senior economist at the French Ministry of Finance, and was previously director of the "Household Taxation" programme at the *Institut des politiques publiques* (IPP). Brice holds a Ph.D. in Economics from Paris School of Economics. He has conducted public policy evaluations related to the structure of social benefits, unemployment allowances, capital taxation, and local taxation. He was also, during eight years, the head of the development of the tax and benefit micro-simulation model TAXIPP, which aims at simulating the French redistributive system using different administrative micro-data sources.

Kamel Gadouche began his career at INSEE to lead several IT projects after studying statistics and IT engineering at ENSAE (CGSA). He joined in 2006 the *Groupe des écoles nationales d'économie et statistique* as IT director. His teaching areas include information security, object programming and conceptual information modeling. In recent years, Kamel Gadouche has led a project to set up the *Centre d'Accès Sécurisé aux Données* (CASD), of which he is now Director.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 12
Quick and Easy, But Decentralised: Access to Administrative Data in Germany



Hanna Brenzel, Toni Schell, Tatjana Mika, and Stefanie Wolter

Abstract This chapter provides a brief overview of Germany's three largest providers of administrative data in the social sciences. The country has a long tradition of providing administrative data for research in a standardised form via research data centres. The data offered are rich in content and are partly available over long spans of time. Access to the data is transparent, fast and inexpensive. Innovations, such as online tools and remote access, make the data even more accessible to researchers and not just those in Germany. However, the infrastructure for data access is decentralised and linking data across institutions involves either lengthy administrative processes or is prohibited by law. Breaking down these legal barriers is a challenge on which the research data centres are currently working together.

1 Introduction

Access to individual data is very fragmented in Germany. There is no "one-stop-shop" in Germany, unlike other European countries (such as, e.g., Austria), but instead there is a large number of accredited Research Data Centres (RDCs) that provide subject-specific individual data for science and research. On the one hand, this occasionally makes it difficult for researchers to find out about the data that is out there and

H. Brenzel
Research Data Centre of the Federal Statistical Office, Wiesbaden, Germany
e-mail: Hanna.Brenzel@destatis.de

T. Schell
Research Data Centre of the Statistical Offices of the Federal States, Duesseldorf, Germany
e-mail: Toni.Schell@it.nrw.de

T. Mika
German Pension Insurance, Berlin, Germany
e-mail: tatjana.mika@drv-bund.de

S. Wolter (✉)
Institute for Employment Research, Nuremberg, Germany
e-mail: Stefanie.Wolter@iab.de

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_12

concerning the requirements for data access. On the other hand, though, the single RDCs are experts on their data. This chapter aims to provide a general overview for those researchers interested in working with administrative microdata from Germany. It also serves as an overview for policy makers on ways to improve data access and, hence, evidence on effectiveness of policies.

The Federal Statistical Office was the first official data producer in Germany to set up an RDC (Statistisches Bundesamt—RDC-Bund) in 2001. The statistical offices of the federal states (Statistisches Landesamt—RDC-Länder), the Federal Employment Agency (Bundesagentur für Arbeit—RDC-BA/IAB) and the German Pension Insurance (Deutsche Rentenversicherung—RDC-RV) followed the example set by the Federal Statistical Office and established their own research data centres. Now, some 20 years later, there are 41 research data centres that are accredited by the German Data Forum (Rat für Sozial-und WirtschaftsDaten—RatSWD) (see Brenzel and Zwick, 2022). The diverse range of data offered by the RDCs includes data from the various disciplines of social, behavioural, educational and economic sciences as well as geographic and spatial data, financial data, statistical data from the federal and state governments and register and social security data. It would go beyond the scope of this chapter to present all of the RDCs that offer administrative data in Germany.[1] For this reason, we have limited ourselves to three of the largest RDCs in Germany and that already cover a large part of administrative data: RDC-RV, RDC-BA/IAB[2] and RDC-Bund und Länder.[3]

We define administrative data as data that is generated within administrative processes and that is either transmitted to the relevant public authority or forwarded from the public authority to a public entity. In order to provide a comprehensive overview of these RDCs, not only are the administrative data of these RDCs are presented, but a complete overview of their product portfolio is also provided. What they all share in common is that the requirements and prerequisites for data access are regulated by law. Access must be requested, is subject to certain conditions and must be contractually agreed. Conditions include confidentiality, restriction to a specific user or group of users, restriction to a specific purpose, and temporal restriction of access.

The remainder of the chapter is organised as follows: Section 2 covers the legal backgrounds of Germany's data access. Section 3 provides an overview of the standard data provided by the RDCs. In Sect. 4, we summarise the similarities and differences in data access, which is followed by sect. 5 which briefly covers statistical disclosure and the consequences for non-standard data products. Finally, section 6

[1] For example, the Deutsche Bundesbank Research Data and Service Centre (RDSC Bundesbank) or the newly established RDC of the Federal Office for Migration and Refugees (BAMF-FDZ) also provide administrative data. An overview of all RDCs and their portfolio can be found on the KonsortSWD's website https://www.konsortswd.de/en/services/research/all-datacentres/

[2] The RDC-BA/IAB is the research data centre of the Federal Employment Agency and is organisationally separated at the Institute for Employment Research.

[3] The RDC-Bund and the RDC-Länder are organisationally separated, but still act together as RDC-Bund und Länder.

provides an outlook on the frontiers of data access in Germany and sect. 7 concludes this work.

2 Legal Backgrounds

This section briefly covers the legal background of data offers and access in Germany and aims to better understand the differences in data access between the RDCs. The main legal sources that comprehensively regulate data access in Germany are the German Social Code (SGB) and the Federal Statistics Law (BStatG).

The RDC-BA/IAB provides data on individuals and institutions that are subject to data protection. The legal basis for data access is mainly § 67 SGB X. Various technical and organisational measures, such as an isolated computing environment, different data access modalities with regard to data anonymisation, contractual regulations and disclosure control allow us to classify the data as anonymised, in accordance with § 282 (7) SGB III. In order to gain access to this data, it is first necessary to comply with a series of strict criteria relating to the topics addressed, the duration of the project and the researchers who are eligible to apply. There are also strict limitations on the merging of any additional information with the data that is being requested. Furthermore, any attempt to de-anonymise the data is prohibited.

The RDC-RV also operates under the German Social Code, SGB X. Data from the pension insurance records are collected as statistical data from the 16 independent legal bodies, the regional and federal pension insurances, which grant pension and rehabilitation according to the same legal framework, the SGB VI. This data collection is already anonymised to a certain degree. The RDC-RV improves the degree of anonymity and offers data in an isolated computing environment as well as for the use at workplaces in both research institutions and universities.

The legal foundation of the RDC-Bund's services is based on is the Federal Statistics Law (Bundesstatistikgesetz—BstatG). For the purpose of scientific research, BStatG permits the provision of microdata, provided that allocating them to the respondents or statistical units "requires unreasonable effort in terms of time, cost and manpower." This presupposes the data's de-facto anonymity. "Within specially protected areas of the Federal Statistical Office and the statistical offices of the Länder" access to formally anonymised data may be granted under certain circumstance. There are technical, organisational and contractual measures that must be adhered to, just as with the RDC-BA/IAB. The use of official microdata is regulated by law, committed to confidentiality, bound to a specific user or group of users, bound to a specific purpose, temporary, contractually agreed and is liable for costs.[4]

[4] For a detailed description see https://www.forschungsdatenzentrum.de/en/terms-use

3 Overview of the Data Offer

This section gives an overview of the RDCs' main data sources and the data products that are derived therefrom. The picture is incomplete as we do not cover survey data. Readers interested in understanding the data sources and products more deeply can find information on the webpages provided in Table 12.1.

3.1 The Research Data Centre of Federal Employment Agency (RDC-BA/IAB)

The RDC-BA/IAB provides administrative microdata and survey data for scientific research. In this section, we focus on administrative microdata only. These data come from two main sources. The first source is the notification to the social security system. Every employer must submit a notification on all employees at least once a year or whenever there is a change in the employment contract that is relevant for the calculation of a person's entitlement to contributions.

The second source is process-produced information. This comes from the systems of the German employment agencies and job centres and contains data about unemployment, job seekers and participation in active labour market programmes. These data sources are then transmitted to the IAB to be transformed into research data and are compiled into Integrated Employment Biographies (IEB). In the end, the IEB includes the universe of all individuals covered by social security and all of the registered unemployed, job seekers and participants in labour market programmes from 1975 onwards and is updated annually. Figure 1 gives an overview of data flows from the source to the RDC-BA/IAB:

(Inter)national researchers have access to standard samples that have been drawn from the IEB and its sources; these are then prepared and documented by the RDC-BA/IAB.

The standard research data include administrative data at the individual and establishment level and combinations of survey and administrative data. The central data sets are: 1) The Sample of Integrated Labour Market Biographies (SIAB), a two per cent random sample drawn from IEB. These data make it possible to track the employment status of a person to the precise day. 2) The Establishment History Panel (BHP) contains cross sectional establishment datasets dating back to 1975 for West Germany and 1992 for East Germany. Every cross section contains all of the establishments in Germany that are covered by the IAB Employment History (BeH) as of June 30th. A 50 per cent sample is available to external researchers. These administrative datasets are partly enriched with survey data on the individual and establishment levels. Moreover, RDC-IAB also offers external surveys that are linked with administrative data (for an overview, see Müller & Wolter, 2020; Schmucker et al., 2023).

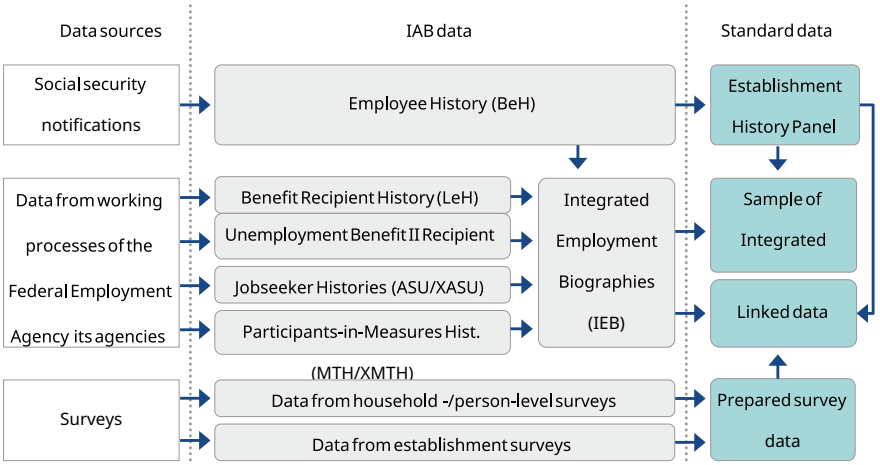


Fig. 12.1 Overview of data sources and products of the RDC-BA/IAB. *Source* own illustration

3.2 The Research Data Centres of Official Statistics (RDC-Bund Und Länder)

The Research Data Centres of official statistics consist of two RDC that are organisationally separated: the Research Data Centre of the Federal Statistical Office (RDC-Bund) and the Research Data Centre of the Statistical Offices of the Federal States (RDC-Länder). Both Research Data Centres cooperate closely in order to meet the growing data needs of empirically minded scientists. They offer a coordinated range of data and services for the scientific use of official statistics microdata. Together, the two RDC offer access to more than 3,000 different data products for over 90 statistics for scientific use via different ways of access. The service covers the following areas: 1) finance, legislation, taxes, 2) economy, 3) social issues, 4) company data and 5) agriculture, energy and environment. The statistics provided by the RDC are not based on a uniform data source, but are instead drawn from a large number of different sources and are collected in a wide variety of ways.

The most widely used, and probably best-known, data set is the microcensus which is an official statistical survey carried out annually since 1957 using a representative sample of one per cent of German households—however, it does not fit in the definition of administrative data. This is followed by a series of economic statistics that can be summarised under the heading "Official Company Data for Germany" (AFiD). AFiD is a modular system and consists of different panel data and modules of economic and environmental statistics that have been both collected and generated by the statistical offices of the Federation and by the federal states. It enables the flexible and individual consolidation of individual datasets. The panel data and modules contain key figures on units from specific economic sectors. The AFiD panel and AFiD modules are usually linked via identifiers of the legal units

(usually the company number) and branches/establishments (usually the establishment number) and can be carried out by the researchers themselves. For example, the AFiD panel "industrial establishments and enterprises" includes units from the mining and manufacturing sectors. The AFiD modules supplement the panel data with information about specific topics in order to answer specific research questions. For example, the AFiD module "environmental protection investment" covers the scope, structure and development of investment in environmental protection. By linking the data, at the company or enterprise level, environmental protection investments made in the commercial sector can be analysed with regard to numerous economic variables. One new AFiD panel concerns the newly established foreign trade statistic. In foreign trade statistics, a distinction is made between extra- and intra-trade statistics. The extra-trade statistics record Germany's cross-border trade in goods with so-called third countries (non-EU states). Intra-trade statistics record Germany's cross-border trade in goods with the other member states of the European Union (EU). The extra-trade statistics are collected on the basis of data from customs declarations and are, therefore, designed as secondary statistics.[5] The intra-trade statistics are a primary survey based on direct reports submitted to the Federal Statistical Office by companies subject to the reporting obligation. Within the area of health, the RDC provide, for example, the Diagnosis-Related Groups Statistic (DRG), which is also a secondary statistic. It covers all of the fully stationary hospital cases in Germany that were accounted for by case rates and on an annual basis. The data is transmitted from the Institute for the Hospital Fee System ("Institut für das Entgeltsystem im Krankenhaus," InEK) to the Federal Statistical Office. These are only a few examples of the wide range of data supply offered by the RDC. However, the aforementioned examples clearly show the diversity of the data provided.

3.3 The Research Data Centre of the Pension Insurance (RDC-RV)

Data from the 16 pension insurance institutions have been collected as part of a coordinated yearly process. The source is the pension insurance accounts that include relevant life-course information about employment, raising children, unemployment, sickness pay and other periods of importance for current or future pension payments. The accounts also include pension benefits and granted medical and professional rehabilitation in the case of severe illness. The data collected for statistical purposes are centred around major interest points of social policy and are collected separately. These themed collections include: accrued pension entitlements for financial planning purposes, actual benefits paid for pensions and rehabilitation and the end of pension payments, which is in most cases death. These focal points of statistical analysis are regulated according to a government regulation, the RSVwV. However,

[5] Secondary statistics are a type of statistic/survey that essentially consist of the transfer of data that was not originally collected for statistical purposes, but for administrative purpose.

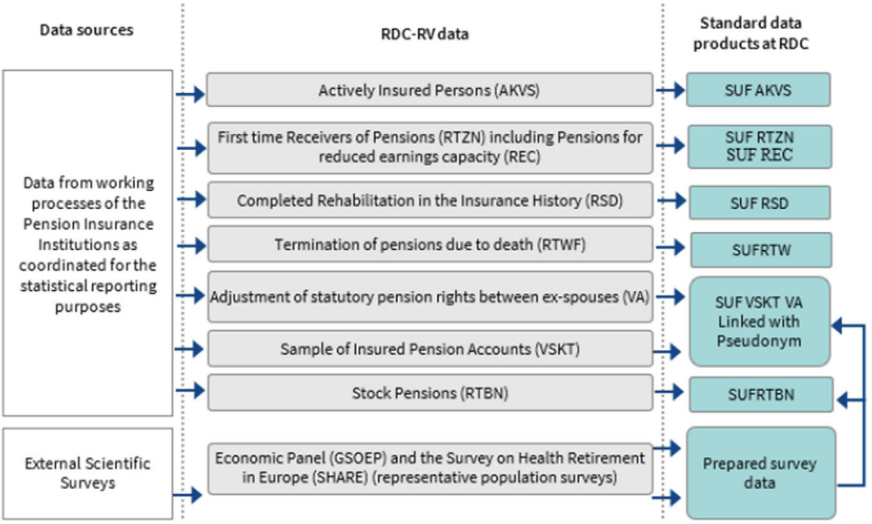


Fig. 12.2 Overview of data sources and products of the RDC-RV. *Source* own illustration, SUF—Scientific Use File

a coordinated process of pseudonymisation makes the data (originally selected separately) linkable for all years since 2011. The resulting linkage options are regularly used for scientific purposes because scientific projects often include a wide array of variables to follow new leads.

Researchers get full access to the data collected for statistical purposes by the pension insurance institutions. Furthermore, scientific surveys like the German Socio-Economic Panel, SOEP-RV (Lüthen et al., 2022), or the Survey on Health Aging and Retirement, SHARE-RV (Börsch-Supan et al., 2020) are linked with data from the pensions insurance accounts on a regular basis. These enriched data give the opportunity to research the whole population on a representative basis and to get a more complete picture of the social situation of insured people and pensioners. Figure 2 gives an overview over the data sources and the derived products.

4 Data Access

Data access is not standardised throughout Germany. This section provides a summary in order to get an overview of key facts about data access. All three RDCs offer a website through which researchers can find out more about the data available and the conditions of access in German and English. The RDCs' employees also offer telephone counselling and advice about data access via email. Applications can be submitted either online or by email to the RDCs. All data providers offer different

access modes that depend on the sensitivity of the data. The more confidential the data, the more restricted the access possibilities become.

The requirements for data access differ slightly between the institutions, particularly with regard to the restriction of research topics. While there are no restrictions with regard to research topics for applicants to the RDC-Bund und Länder, greater restrictions in terms of research topics apply for data access at both RDC-BA/IAB and RDC-RV. The data access is free of charge for data at RDC-BA and RDC-RV and small fees apply at RDC-Bund and Länder. Table 12.1 gives a comprehensive overview of key points of data access.

Table 12.1 Data access in Germany

	RDC-Bund und Länder	RDC-BA/IAB	RDC-RV
Information on data offer	https://www.forschungsdatenzentrum.de/en	https://fdz.iab.de/en	www.fdz-rv.de
Data portfolio	The coordinated range of data and services includes over 90 statistics on economic, financial, environmental, social and other topics. Social statistics: Population, education, health, households, child and youth welfare and other social statistics. Agricultural, energy and environmental statistics Finance, legal and tax statistics Economic statistics: Construction and housing, services, manufacturing and other economic statistics Company data (AFiD)	Administrative labour market data on individual and establishment level from apprenticeship training to retirement Survey data on individuals, households and establishments covering (non-) employment, education, workplace relations and migration among others Linked survey and administrative data	Administrative data on insurance and pension entitlements Data about old-age insurance Thematic data on demographics and mortality, insured pension inflow and the reduction in earnings capacity Data on medical and vocational rehabilitation
Application	The request can either be filed via an online form or sent via postal services or email.	Online word document that has to be emailed/printed.	The request is filed online.
Contract	Contract needs either original signature or certified digital signature.	Contract needs either original signature or certified digital signature.	Contract needs either original signature or certified digital signature.
English language application/ contract available	Yes/no	Yes/yes	Yes/yes (on request)

(continued)

(continued)

	RDC-Bund und Länder	RDC-BA/IAB	RDC-RV
Eligible to apply	Universities and research institutions that conduct independent scientific research. Universities that are recognised within the ANABIN-database.	Universities and research institutions that conduct independent scientific research. Institutions from EU countries and countries that are recognised as having an equal standard of data protection.	Universities and research institutions that conduct independent scientific research. Institutions from EU countries and countries that are recognised as having an equal standard of data protection.
Restrictions on topics of scientific research	No restriction.	Topics within the field of labour market research.	Topics within the field of social security.
Costs	A fee is charged. The amount of the fee depends on the number of statistics required, survey years and ways of data access, and varies according to the different levels of effort involved in each access method. Extra costs for project-specific processing are incurred.	Free of charge for standard products	Free of charge for standard products
Avg. time from application to access	2–4 weeks	2–4 weeks	1 week for standard SUF, 3 weeks for on-site access at the RDC
Documentation	Data and methodological reports, test data, codebooks and frequencies. In German (English available for microcensus, DRG and EU-SILC)	Data and methodological reports, test data, codebooks and frequencies. In German and English	Data and methodological reports and codebooks. Frequencies on the statistical information website. German and some codebooks in English.

(continued)

(continued)

	RDC-Bund und Länder	RDC-BA/IAB	RDC-RV
Data access modes	Depends upon the sensitivity of the data: Off-site as scientific use file (SUF) or public use file (PUF)/ campus file (CF) On-site via remote execution or within safe Centre.	Depends upon the sensitivity of the data: Off-site as scientific use file (SUF) or campus file On-site at 20 sites in EU, UK, USA and Canada and via remote execution	Depends upon the sensitivity of the data: Off-site as scientific use file (SUF) On-site and via remote execution.

5 Data Protection

This section briefly gives an overview of data protection and the statistical confidentiality of the microdata offered by the RDCs. A central element of data access in Germany is the anonymity of natural and legal individuals that are part of these data products. Identifiers such as names, addresses or personal identifiers are either deleted or replaced by pseudo identifiers. Information that is particularly sensitive is either aggregated or not offered among RDC data products. Moreover, special rules for sensitive data also apply. They can only be accessed from workspaces within the RDCs that provide isolated computers without internet access. Research output from sensitive data also undergoes a security check. These measures are taken in order to prevent the deanonymisation of individuals or firms.

The rules described about data security make clear that it is not possible to link research data of the RDCs to other data sources on the micro level. However, linking statistical data (e.g., on a regional or industrial level) is possible. It is essential for some research projects to link data sources on a micro level. The legal regulation of the institutions and the GDPR do not prohibit such kinds of projects, but do set out a strict and narrow legal framework. Even if all requirements are met, information that identifies individuals is not disclosed to external researchers, but instead to employees of the RDC who will link data and provide anonymised data to the external researcher thereafter. Individual data samples and linkages are time-consuming and expensive and are not the task of the RDCs in Germany.

6 Frontiers in Data Access

Data access is not static. The RDCs collaborate in order to foster innovation and to improve services for researchers. This section summarises some of the frontiers.

The decentralised access to data is driven by different legal bases, making it even more difficult to link data across institutions. Data providers and researchers

are working together in order to overcome some of these barriers. Currently, there exists no overall procedure by which to merge data between different institutions or RDCs. Progress, however, is slow and mainly possible for individual projects, not on a common basis. Linking data across institutions is a time-consuming and partly costly way by which to access data. An easier way to link data is possible, provided that the linkage involves only one administrative data provider and other non-administrative data. The case-by-case approach is mainly applied where rules and procedures have to be negotiated and a legal basis has to be proofed for each individual request.

Some more generalised linkage possibilities have been created, despite this case-by-case approach. Best practice concerns the linking project between the RDSC Bundesbank and RDC-Bund (see Boddin et al., (2024)). It is now possible, for the first ever time, to link research data from the research data centres of the Federal Statistical Office and the Statistical Offices of the Federal States (FDZ) with those of the Research Data and Service Centre (FDSZ) of the Bundesbank. The linked data can be accessed. Separate rules concerning both the use of and joint data governance have been created for this express purpose. Another cooperation between RDC-BA/IAB and RDC-RV linked employment and pension data. The linking process and subsequent quality assurance were very complex and revealed the hurdles of this process. Nevertheless, cooperation involving only one state institution can pave the way to simplifying cooperation between state institutions at a later date. These cooperations show what is feasible and what might even help to dispel fears about a reduction in data protection.

A second frontier concerns the way in which data is accessed. At present, users have to travel to data safe rooms in order to access more sensitive data, which increases indirect costs for data access and excludes some potential users, or uses remote execution which is a user-unfriendly and time-consuming way by which to access data. The data providers in Germany are increasingly making their way towards establishing remote desktop access. This data access method is seen by many researchers as the ideal way through which to analyse data and has long been established internationally as a way to access data. A remote desktop will enable researchers to access data from their offices through a secure connection to the research data centre. Data transmission in a physical sense does not take place. It is only possible to view and analyse the data within the secure system. The RDC-BA/IAB, as well as the RDC-Bund and Länder,[6] are currently in a piloting phase for remote desktop. This will only be accessible for researchers in Germany initially. The new way of access will not replace existing and established data access methods (e.g., remote execution), but will instead complement it. Furthermore, it will not be possible to replace on-site stays completely, given that not all data fulfils the requirements for remote desktop and non-EU countries will not be allowed to access data through this direct channel.

Last, but not least, some research data centres also cooperate internationally to harmonise data access and to lower barriers for data access. The IDAN Network (https://idan.network/) is a grassroots movement of four European research data

[6] See for more information Brenzel et al. 2023.

centres that cooperate to promote remote cross-access between countries to secured data.

All of these initiatives operate within the narrow and strict framework of the German interpretation of data protection laws.

7 Summary

This section provides a comprehensive overview of the legal backgrounds, data sources and products offered by the research data centres RDC-BA/IAB, RDC-Bund and Länder and RDC-RV. The three main providers of administrative data for social research in Germany handle a significant number of the overall projects in this field from all of the RDCs registered at Rat SWD. Despite different legal backgrounds for data provision in RDC-IAB, RDC-Bund and Länder and RDC-RV, key structures and requirements remain comparable. Each RDC is specialised on a data portfolio and can advise potential researchers about its data. Details about the application process, however, are different. Overall, data access in Germany is transparent, incurs a low cost and is fair. This also applies to research data that is available through the RDCs. The need for special samples or data that is linked across different RDCs marks the boundaries of data access. The progress made to overcome these boundaries is slow and are often hindered by strong legal ties. Nevertheless, pioneering projects show that linking data across different institutions is not only possible, but can also fulfil the high data protection requirements.

References

Boddin, D., Kruse, H. W., Merkle, H., Walter, S., & Zapf, B. (2024). Combined international trade and investment data "CITID"—A new innovative and comprehensive data landscape for Germany. *Jahrbücher für Nationalökonomie und Statistik, 245*(3). https://doi.org/10.1515/jbnst-2024-0024

Börsch-Supan, A. C., Friedel, S., Herold, I., Korbmacher, J., & Mika, T. (2020). SHARE-RV: Linked data to study aging in Germany. *Jahrbücher für Nationalökonomie und Statistik, 240*(1), 121–132. https://doi.org/10.1515/jbnst-2018-0034

Brenzel, H., Cramer, K., Güttgemanns, V., & Mathes, M. (2023). *Remote Access for Scientific Use Files—a New Pathway for German Official Statistics Microdata Access*, https://unece.org/sites/default/files/2023-08/SDC2023_S1_1_Germany_Brenzel_D_0_0.pdf

Brenzel, H., & Zwick, M. (2022). *An informational infrastructure has emerged in Germany—The research data Centre of the Federal Statistical Office (WISTA 6).* Federal Statistical Office.

Lüthen, H., Schröder, C., Grabka, M., Goebel, J., Mika, T., Brüggmann, D., Ellert, S., & Penz, H. (2022). SOEP-RV: Linking German socio-economic panel data to pension records. *Jahrbücher für Nationalökonomie und Statistik, 242*(2), 291–307. https://doi.org/10.1515/jbnst-2021-0020

Müller, D., & Wolter, S. (2020). German labour market data—Data provision and access for the international scientific community. *German Economic Review, 21*(3), 313–333. https://doi.org/10.1515/ger-2019-0127

Schmucker, A., Stefan, S., & vom Berge, P. (2023). *Stichprobe der Integrierten Arbeitsmarktbiografien (SIAB) 1975-2021* (FDZ-Datenreport 02/2023 (en)). *Research Data Center of the Federal Employment Agency in the Institute for Employment Research.* https://doi.org/10.5164/IAB.FDZD.2302.de.v1

Hanna Brenzel studied international economics, with a focus on Central and Eastern Europe, in Regensburg. From 2012 to 2018, she was a researcher at the Institute for Employment Research in Nuremberg. In 2016, she managed a labour market integration project for refugees. She completed her PhD in 2018 at Otto Friedrich University in Bamberg on the topic of the "Economic Integration of Migrants." Since 2019, Hanna Brenzel has been Head of the Unit "Methods of Data Analysis" at the Federal Statistical Office and, since August 2021, also the Head of the Research Data Centre of the Federal Statistical Office. Since 2022, Hanna Brenzel has been the Principal Investigator (PI) of the research group FOR 2559 "Cross-sectoral small-scale microsimulation model" (MikroSim), which was funded by the German Research Foundation (DFG).

Toni Schell studied social sciences with a focus on quantitative empirical social research in Düsseldorf. He completed his doctorate on the topic of "Personalization trends in voting decisions" at Heinrich Heine University Düsseldorf in 2019. From 2012 to 2015, he held a scholarship at the "Linkage in democracy" Graduate school at the University of Düsseldorf. From 2015 to 2019, he worked as a research assistant at the Fern Universität in Hagen. Since 2019, he has been a consultant at the Research Data Centre of the Statistical Offices of the Federal States.

Tatjana Mika studied law and sociology at the Freie Universität Berlin, University of Padova and the New School for Social Research. She received her first state examination in law from the Land Berlin as well as the second state examination in law. Tatjana Mika has been a lecturer at the Humboldt University for Sociology of Law at the Faculty of Law and is a researcher at the Humboldt University at the "International Social Justice Project" at the Department of Social Sciences. Since 2002, she has been working at the research data centre of the German Pension Insurance and since 2011 has been working as head of the RDC-RV. Under her leadership, the RDC-RV established scientific use files for linked survey and record data, especially SHARE-RV and SOEP-RV. Tatjana Mika also conducts her own research, which has been published in international peer-reviewed journals.

Stefanie Wolter studied European Studies at the Chemnitz University of Technology and International Economics at Regensburg University. She received her doctorate from Julius-Maximilian-University Würzburg in 2019. Stefanie Wolter has been a researcher at the Research Data Centre of the Federal Employment Agency since 2011. She is involved in developing new data sources. Among other things, she is leading a project that conducts surveys of establishments and employees and then links this data with administrative labour market data. This has resulted in an internationally unique data set, the Linked Personnel Panel. Under her leadership, establishment data was linked to the enterprise level and was supplemented with balance sheets and shareholder information. Stefanie Wolter also conducts her own research and is published in international peer-reviewed journals.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 13
The Limited Use of Administrative Data in Greece



Daphne Nicolitsas

Abstract In Greece, as in most EU countries, administrative data are collected for several purposes: tax reasons, social security administration, school records, population changes and unemployment registers, to name but a few. The advantages of using administrative data, compared to survey data, cannot be overemphasised (see, inter alia, OECD, 2020; Røed and Raaum, 2003). The collection of administrative data was, until recently, a lose-lose situation in Greece: on the one hand, those having to report them complained of the bureaucratic burden while researchers who could have benefited from using data were not granted access to them on the other. Although the use of administrative data in Greece has now increased, the main user is the Hellenic Statistical Authority (ELSTAT), with researchers benefiting in an ad hoc, mostly indirect fashion.

1 Introduction

This chapter reports on both the availability and use of administrative data in Greece. The chapter is organised as follows: Section 2 documents the best-known administrative databases for Greece and Section 3 reports on the administrative databases used by ELSTAT. Section 4 presents examples for which available administrative data would be useful and also includes examples of such data being used. Finally, Section 5 summarises and concludes.

D. Nicolitsas (✉)
Athens University of Economics and Business, Athens, Greece
e-mail: nikolitsa@aueb.gr

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_13

2 Available Administrative Databases for Greece

2.1 Population Statistics

Birth and death registration data are the main population statistics. In what follows, we describe the data collected for these two events. Births can now be registered online by parents before leaving the maternity ward and data are, thus, available immediately, contain fewer errors and are up to date.

Birth registration data includes information about the parents, such as identifying information (surname, name and tax identification number [TIN]); religion; nationality; occupation; social security fund; permanent address; parents' education; parents' marital status and wedding date if the child was born in wedlock. Information concerning the birth concerns whether the birth was part of multiple births; pregnancy duration; c-section or not; whether a doctor, midwife and/or nurse was present at birth; place and area where the birth took place and date of birth. There is also information about the child, such as sex and weight, as well as on the number of children to which the mother has previously given birth.

Death registration data includes information about the deceased, such as identifying information (surname, name, TIN and social security number), religion, nationality, occupation, permanent address, marital status and education. Information is also included regarding the cause and location of death, as well as identifying information about the doctor testifying to the death.

2.2 Business Registry Data

A comprehensive registry of all companies that are active in Greece has also been established. The General Commercial Register, known by the acronym GEMI,[1] includes basic demographic information on all enterprises that are active in Greece. More specifically, the registry includes *Company identifying information*: company name, identification number (ID) and European Union ID (EUID), legal entity type, sector of economic activity, date of establishment, date of closure and foreign or domestic subsidiary. The database also includes *Demographics of natural persons involved in company management* (age, gender, position in company and nationality) and the *latest annual financial data on companies.*

[1] The establishment and operation of GEMI was provided for in L. 3419/2005. GEMI came into operation in April 2011. The law governing its operation is L. 4919/2022.

2.3 Labour Market Statistics

The *Ergani database* covers administrative data on the labour market, providing both a monthly snapshot and information about the flows in and out of the labour market.[2] Most of the data now collected through the Ergani database had been collated from hand-written documents for decades. It was not until 2013 that the Ministry of Labour and Social Security created a continuously updated information system to measure the flows of employment in the economy's private sector in real-time.[3] The Ergani database includes the information presented below.

Detailed information on workforce: A monthly snapshot of all employees includes each employee's recorded social security number, work experience, occupation, daily and weekly hours of work, remuneration (monthly, daily and hourly), marital status and number of children.

Hiring notice: Employers must notify the Public Employment Service (DYPA) when hiring an individual. The notice includes the employer's identity (company name, legal entity type, TIN, activity, location and employer's association membership). It also includes the employee's identity (name, sex, age, social security number, TIN, nationality, marital status, level of education and training). Details about the job for which the individual was hired are also provided: start date, weekly working hours, employment status (full-time, part-time and alternating), contract type (indefinite, fixed-term), occupation and remuneration.

Termination notice: Much in the same way as when an employee is hired, employers must also notify the DYPA of any employment termination. In addition to the information provided when hiring the individual, the termination notice includes details regarding the termination of the employment relationship: voluntary resignation, dismissal, collective dismissal, notification period, final salary and dismissal compensation.

It is unclear whether all of the Ergani database's fields are filled in or whether there are gaps in the information recorded (e.g., membership of employers' associations may not be recorded).[4] The Ergani database is *not* available through ELSTAT.

Greece's Public Employment Service, known as DYPA, also has administrative labour market-related data from the individuals registered with its services. More specifically, DYPA has information about the demographic (age, gender, family status) and productive (education, work experience) characteristics of the individuals registered with it, as well as information on registrants' participation in any Active

[2] According to Greek mythology, Ergani ("The Industrious") is an epithet given to the goddess Athena as the patroness of labour, arts and sciences.

[3] Law 4152/2013, Subsection IA.3.

[4] This is likely the case as employers' associations are not forthcoming with information on their members even though L. 4635/2019 (Art. 54) provides for the establishment, at the Ministry of Labour and Social Security, of a General Registry of Employers' Organisations, in which all employers' organisations are compulsorily registered and which contains details on inter alia the number of members of the employers' organisation which participated in the elections for the association's governing body.

Labour Market Programmes (ALMPs) run by DYPA and the duration of unemployment spell. These data were recently provided to the OECD who, after linking them with the Ergani data, was able to study the efficacy of different ALMP programmes run by DYPA. There is no prescribed procedure for obtaining the DYPA data.

2.4 Social Security Statistics (SSS)

Employers are obliged to report the following information for each employee monthly: demographic information (date of birth, TIN, Social Security Number), productive information (occupation), remuneration and working hours and social security contributions paid. Information about the employer's economic activity and social security contributions is also recorded. As a result, the Unified Social Security Fund (EFKA) has a database that includes information about Greece's population of both employees and employers.

2.5 Tax Statistics (TS)

Tax statistics are the only administrative dataset with a prescribed access procedure at the source (i.e., not via ELSTAT). Data are mainly available from the Independent Authority for Public Revenue (AADE) to the state's legal entities, rather than individuals, and a detailed account of what is permitted or forbidden is agreed on between the parties before data is released. ELSTAT is given access, inter alia, to data necessary for national account estimates. According to the provisions of L. 5104/2024 (Article 21), and in order to ensure confidentiality, persons who are officials of the Tax Administration and the Ministry of National Economy and Finance may grant access to individual-level tax and income data for ELSTAT-authorised personnel and to such individuals authorised by ELSTAT. They may also grant aggregate statistical data held by the tax administration, with the obligation to use the data for the purpose for which these were requested within the framework of the competencies of ELSTAT or of the other entities or authorities and respecting confidentiality clauses to ELSTAT-authorised personnel, to other general government entities and to independent regulatory authorities. An application form in Greek is readily available to apply for tax statistics.[5] The management of applications is undertaken on a case-by-case basis through an open procedure. Projects are judged on the basis of the need to use this data and on the fulfilment of data protection rules. Data can be accessed remotely.

[5] http://elib.aade.gr/elib/view?d=/gr/ap/2018/1154/

2.6 Education Statistics

ELSTAT collects statistics on students, teachers and infrastructure at all levels of education. Such information includes total student number per school, total number of teachers/faculty, average student age and total number of students by gender, as reported for every prefecture. These data are aggregates at school level, given that they are collected by schools.

The Ministry of Education, Religious Affairs and Sports maintains a comprehensive administrative database at the student level, including demographic information on the student (age, gender, nationality and area of residence) in all types of schools in Greece and Greek schools operating abroad, school grades in all subjects and information about students' parents (occupation). It also includes information about each faculty member (length of time at each school, work experience, type of employment relationship, work hours and absences). To the best of our knowledge, this database, known as *Myschool*, has not yet been released to any researchers. Furthermore, the Ministry of Education maintains a database with the grades of all of the students taking university entrance exams, their preferences for the departments that they wish to enter and the outcome (*Panhellenic exam grades database*). The Myschool and Panhellenic exam grades databases could easily be linked by using existing unique student identifier numbers.

2.7 Other Statistics

Finally, administrative data from the justice system, such as rulings reached in each trial in the Civil Court, Court of Appeal and Court of First Instance, as well as by the Judge-Rapporteur are reported and distributed through ELSTAT. The statistics only record limited information about the particulars of those involved.

3 Use of Administrative Data by the Hellenic Statistical Authority (ELSTAT)

ELSTAT is the main research-oriented user of the administrative data that is collected in Greece, which it draws on to perform its surveys. More specifically, ELSTAT uses the business register described in 2.2 above to compile samples for surveys producing Structural Business Statistics (SBS) and Short-Term Business Statistics (STS), for surveys used for foreign affiliates statistics (FATS), for surveys on ICT usage by enterprises and e-commerce, the Prodcom survey and the trade by enterprise characteristics (TEC) survey (ELSTAT, 2016). The use of the national business registry by ELSTAT is, in any case, provided for by European Parliament and Council Regulation (EU) 2019/2152.

ELSTAT is also the distributor of administrative data in certain instances. The authority has a prescribed procedure for granting data to researchers. This involves, inter alia, the submission of an application consisting of a research proposal to illustrate the data's intended use and a contract between ELSTAT and the user concerning storage and confidentiality. Submission of applications is not restricted to those affiliated with research institutions, although it is unclear whether the data would be released for commercial purposes. The research proposal must include the following information: legitimate purpose of the research; the research project, research organisation commissioning the research or requesting access to the research; individuals authorised by the research organisation/researcher who will have access to the data; form in which results will be published and time required to carry out the research. The research proposal goes before the Statistical Confidentiality Committee, which assesses whether granting the data would violate confidentiality in any way.[6] Thus, each application is assessed separately; there is also no general rule as to whether the data will be provided or not. The Statistical Confidentiality Committee comprises ELSTAT employees and experts in statistical methodology, statistical surveys, informatics and in the field of the dissemination of statistical information. If the request for data is approved, then both ELSTAT and the researcher sign a contract concerning the confidentiality of the statistical data and sanctions in case of breach-related obligations by the researcher. The Confidentiality Committee convenes in the event that there are issues, rather than at a regular, pre-announced frequency. The data provided are anonymised; thus, linking them with other databases is impossible. Data are provided to the researcher so that she can work on her own premises and no provision for remote access to more detailed data is made. The contract signed with ELSTAT also asks researchers that they use the data exclusively for scientific ends, thereby prohibiting the use of the data for any other purpose, including financial or other personal or third-party benefit, as well as any political, collective or otherwise defined purpose or any other purpose of any kind, including the use of data for the imposition of taxes, duties, fees and related charges, or before tax, judicial, police, military or other administrative authorities.

ELSTAT grants access both to the surveys it conducts itself and to other administrative databases that it receives, such as the population statistics and justice statistics described in Section 2. Documentation by ELSTAT is available for its own surveys, as well as for any administrative data provided. Documentation quality differs according to the database's popularity. Documentation is detailed for the flagship Labour Force Survey (LFS), which is provided in Greek and English; for other databases, though, documentation is less detailed, but ELSTAT personnel are willing to provide as much information as possible. However, this implies additional efforts by researchers and comes with concomitant delays.

[6] The provisions for the Statistical Confidential Committee are set out in the Regulations on the Operation and Administration of ELSTAT. The regulations are available in English from the following link https://www.statistics.gr/documents/20181/301069/ELSTAT_REGULATION_%CE%95%CE%9D/66c62d27-43c8-4132-b96f-3b5a33e7d7b5?version=1.0 (accessed 11.06.2024).

4 Uses and Potential Uses of Administrative Data

As mentioned above, ELSTAT is the main user of administrative data in Greece and draws upon them to conduct its surveys. However, the use of raw administrative data would better serve research and evidence-based policymaking, given, inter alia, the declining response rate in surveys. One indicator of the extent to which surveys have lost their appeal is the declining response rate to ELSTAT surveys. Figure 13.1 shows the non-response rate for the LFS, ELSTAT's flagship survey, and the survey recording vacancies in enterprises. Response rates to other surveys, run by individual researchers working at the Bank of Greece, have even lower response rates. See, for example, Nicolitsas (2016) concerning the response rate to a survey on wage- and price-setting practices run by the Bank of Greece and IOBE under the auspices of the ECB as part of the Wage Dynamics Network (WDN) for which the response rate was less than 10%.

Of all of the administrative data described in Section 2, the Ergani data have only been used by research institutes and social partners participating in discussions about the minimum wage. Even then, the Ergani data provided by the Ministry of Labour and Social Security were not, in general, provided at the individual level, but instead as averages of cells by age, education, employment type, etc. A recent OECD report (OECD, 2024), however, has used individual-level data from Ergani to assess the efficacy of wage subsidies and training on job retention or employment probability. The report often alludes to errors in the wages recorded in the past in Ergani that have

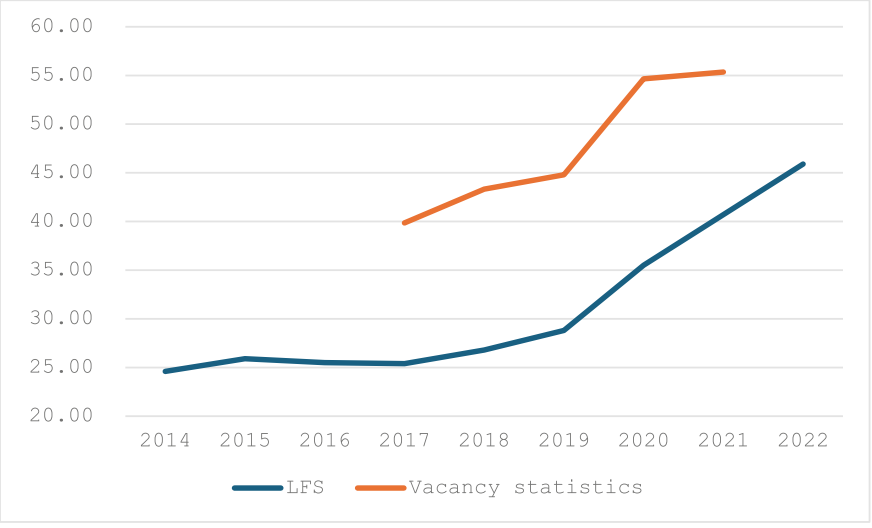


Fig. 13.1 Non-response rates to the LFS survey and vacant positions survey. NB: The non-response rate for the LFS survey in 2021 is missing and the data point above is an interpolation between 2020 and 2022. Source: ELSTAT site

not been corrected.[7] No published procedure exists according to which to obtain the Ergani database.

Conversely, the Social Security Statistics database (SSS) has been used previously in research. The two such investigations that we are aware of included, first, a study by Georgiadis et al. (2020) using a random sample from Unified Social Security Fund (EFKA) records, as described in Section 3 above, to study the impact of the decrease in the minimum wage on adult and youth employment in 2012. Second, a study by Saez et al. (2012) used all records from the SSS, around the time of a policy change, on the maximum social security contributions to be paid. The research question aimed to determine how social security contributions affect labour supply, taking advantage of the policy change for identification purposes.

Access to data is given ad hoc by the SSS. An application must be submitted to the Unified Social Security Fund (EFKA), who evaluate it and decide whether to grant the data. More details about the procedure to be followed, and the criteria on which data are provided, are not available. Anecdotal evidence suggests, however, that the workload of EFKA employees is one of the criteria on which the decision is based. SSS data are not accessible through the ELSTAT procedure.

Several policy issues in the last few years could have benefited from administrative data research results. In 2016, a pension reform introduced a virtually flat "national pension" that was payable to all individuals with at least 15 years of contributions, plus a "contributory" pension, linking contributions over an individual's entire working life to entitlements (European Commission, 2017). Calculations of past pensionable earnings, however, require an adjustment for wage inflation. The initial law provided for an adjustment that was based on average wage inflation. However, no comprehensive wage data series existed, so the law changed so as to provide for an adjustment according to price inflation until wage data became available. Another issue involved the extension of collective wage agreements beyond the parties that negotiated the agreement. Administrative data could also assess whether sectoral and occupational collective agreements should be extended to all firms or employees in a given occupation. The current rule permits an extension, provided that the employers who negotiated the agreement represent at least 51% of the employees in the sector. As far as we are aware, the threshold is calculated without using the Ergani data that include information on a company's (main and secondary) activities and how its employees are occupied, which would prove extremely useful in the case of occupational-level collective wage agreements.

Administrative data could have been used to assess the benefits of reducing social security contributions on employment. Finally, there is a long-running dispute between the current government's former Minister of Education and a well-known public figure and former minister concerning the necessity of publishing the success rate in university entrance exams as a means to assess school efficiency.[8] This is

[7] "In particular, the Ergani data on earnings and employment contain a significant number of implausible and inconsistent values, especially in earlier periods" (OECD, 2024; p. 57).

[8] See, for example, https://www.liberal.gr/epikairotita/o-skotadismos-tis-k-kerameos (in Greek, accessed 11.06.2024).

possible, given the *Myschool and Panhellenic exams databases* described in Section 3 above.

5 Summary and Conclusions

This chapter has described several administrative data that are available means through which to research both the economic and social issues that are of interest to the Greek economy and society. More data are available than the research using such data would lead us to believe. The heavy workload faced by public sector employees, together with concerns regarding the accuracy of the data and confidentiality-related issues, seem to be behind this lag in research output. Recently, some steps have been taken to release administrative data for research purposes. However, there is a risk that this may lead to accusations of unfair and or non-transparent procedures to obtain access to Greek administrative data. As a consequence public discussion about access to administrative data might easily become distracted from the main, over-arching objective, namely the need for evidence-based policy.

References

ELSTAT (2016.), "Use of administrative data for the compilation of structural business statistics", Presentation at the 2016 ELSTAT users' conference, https://www.statistics.gr/el/user-conference-2016 (in Greek).

European Commission. (2017). The ESM stability support programme: Greece, first & second reviews July 2017 background report. *Institutional Paper, 064*. https://ec.europa.eu/info/publications/economic-and-financial-affairs-publications_en

Georgiadis, A., Kaplanis, I., & Monastiriotis, V. (2020). Minimum wages and firm employment: Evidence from a minimum wage reduction in Greece. *Economics Letters, 193*, 8. https://doi.org/10.1016/j.econlet.2020.109255

Nicolitsas, D. (2016). Price setting practices in Greece: Evidence from a small-scale firm-level survey. *Bulletin of Economic Research, 68*(4), 367–382. https://doi.org/10.1111/boer.12058

OECD (2020), *Impact evaluation of LMP through the use of linked administrative data*, Paris: OECD. (https://www.oecd.org/els/emp/Impact_evaluation_of_LMP.pdf, accessed 30.06.2024).

OECD. (2024). Impact evaluation of training and wage subsidies for the unemployed in Greece. *Connecting People with Jobs, OECD Publishing, Paris*. https://doi.org/10.1787/4b908517-en

Røed, K., & Raaum, O. (2003). Administrative registers—Unexplored reservoirs of scientific knowledge? *The Economic Journal, 113*, F258–F281. https://doi.org/10.1111/1468-0297.00134

Saez, E., Matsaganis, M., & Tsakloglou, P. (2012). Earnings determination and taxes: Evidence from a cohort-based payroll tax reform in Greece. *The Quarterly Journal of Economics, 127*(1), 493–533. https://doi.org/10.1093/qje/qjr052

Daphne Nicolitsas is an Assistant Professor at the Department of Economics of the University of Crete. Before that, she worked in economic policy-related jobs and the financial sector (Bank of Greece, Greek Ministry of Economics and Finance, Emporiki Bank, Tudor Proprietary Trading). She has held research posts at the University of Manchester and the University of Oxford. She

holds a BA from the Athens University of Business and Economics, an MPhil from the University of Oxford and a PhD from the University of Manchester. Her main research interests lie in labor economics and industrial organisation, fields in which she has published in international journals.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 14
Rich Data, Strong Laws to Promote Access, Weaknesses and Uncertainties in Implementation: The Case of Hungary



Ágota Scharle

Abstract Facilitated by a favourable legal framework, the secondary use of anonymised administrative data has expanded significantly for applied research during the past decade in Hungary. However, the more recent developments regarding both the access and usage of administrative data have been mixed. There is a new initiative, stemming from the government, that signals increased interest and expertise in using available data to support decision making; alongside new legislation to define rules of access, a new agency was also set up with a mandate to map the available administrative data and to promote secondary usage. However, the new agency's main focus is to serve government decision makers, while data access has become increasingly uncertain for independent researchers and policy analyses.

1 Introduction

The legal framework for accessing anonymised administrative data was relatively favourable in Hungary until recently. This gave rise to an expansion of independent research based on administrative data, whereas secondary analysis to support government decision making remained limited. The re-regulation of data management and accessibility, enacted between 2021 and 2023, might reverse this situation by giving priority to the secondary use of data within the government and by not specifying a clear procedure by which to request linking by independent researchers (for purposes unrelated to government decision making).

This chapter's next section briefly outlines the recent legal and institutional developments concerning access to administrative data for secondary analysis and assesses implications for independent research. Section 3 describes currently available data and the rules of access, while Sections 4 and 5 provide some examples of policy-relevant research based on administrative data in Hungary. The final section considers the main barriers to expanding secondary analysis within the government.

Á. Scharle (✉)
Budapest Institute for Policy Analysis, Budapest, Hungary
e-mail: agota.scharle@budapestinstitute.eu

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_14

2 Recent Changes in the Legal Provisions for Accessing Administrative Data

The legal framework for accessing anonymised administrative data was, until recently, relatively favourable in Hungary. Act CI of 2007 ensured a clear framework for requesting the anonymisation of administrative data, as well as the linking of such data, with some restrictions that served to balance the costs of data management against the public interest of expanding the secondary use of public data. The public agency responsible for both linking and anonymising administrative data (the National Infocommunications Service Company, NISZ) neither actively promoted the secondary use of administrative data nor the anonymised datasets that they had compiled, but provided a decent service to those who submitted a request for linking.

As reported in Scharle (2019), the secondary use of administrative data for policy analysis had remained limited within the government, but has expanded significantly in applied research, fostered by the establishment of a Databank at the Centre for Economic and Regional Studies of the Hungarian Academy of Science.[1] The Databank was not involved in the technical process of linking the data, but has acted as an intermediary between researchers and NISZ (the public agency responsible for the linking and anonymisation of administrative data).

Legal and institutional conditions changed in June 2021 with the enactment of new legislation concerning national data assets (Act XCI of 2021 and government decree 607/2021). The new law repealed Act CI of 2007 and established the National Data Asset Agency (Nemzeti Adatvagyon Ügynökség, NAVÜ) with a mandate to both create and maintain a registry of administrative data and a portal to manage data access. However, the new law did not specify a clear procedure for claiming access to anonymised administrative data; the procedure for linking such data is likewise unclear.[2] These details were provided by another piece of legislation, Act CI of 2023 concerning the system of utilisation of *national data assets* and certain services, which repealed Act XCI of 2021. Act CI of 2023 outlines a gradual process by which to improve access to microdata. Linking, anonymisation and analysis are primarily conducted by NAVÜ staff in the first, current stage. A safe room and remote access are planned to be introduced in the second stage, starting in 2026. The current legislation also lists the provision of microdata for the purposes of analysis (§51) as one of NAVÜ's services, but does not specify who may request this service.

NAVÜ has a much stronger mission to promote the secondary use of administrative data compared to NISZ, whose main focus is on maintaining the IT infrastructure of government offices. The new laws expressly mandate that NAVÜ develop facilities to support users and to systematically map datasets that may be potentially useful for secondary usage. If these developments materialise, then this may significantly improve the visibility and accessibility of administrative data for all. What is still

[1] It should be borne in mind that the Databank was set up by researchers as a grassroots initiative.

[2] For the broader policy context of the recent legislative changes, see Sántha and Sikolya (2017), Büki et al. (2022) and Németh (2021).

uncertain, however, is who or what institutions may make a request for linking datasets.

Another important change is that NAVÜ has been moved under the control of the Central Statistical Office as of February 2024. This reduces the risk of mission drift in NAVÜ and may also strengthen expertise in both data management and analysis. As the Central Statistical Office already has a safe room, their supervision may also speed up the establishment of a safe room in NAVÜ.

Though the current legislation reflects a strong commitment to promoting the secondary use of public data, the delays in creating the new infrastructure for claiming and accessing administrative data cast some doubt on the government's underlying motivations. The official explanations for introducing the new law[3] refer to exogenous motivations primarily, such as that the ranking of Hungary should be improved in the Digital Economy and Society Index (DESI) and the need to implement the European Data Governance Act of 2022.

By June 2024, NAVÜ had already created a searchable registry of administrative datasets (https://kozadatportal.hu/publicdata), which provides meta-data for about 372 datasets that are maintained by various public agencies, but it is only available in Hungarian at present. The registry's information is very basic: it covers contact details and the broad topic, but provides no further details about the time frame, contents or identifiers that may allow linking to other datasets.

3 Administrative Data Available for Research Purposes

Though the current legal provisions and infrastructure do not facilitate access to most recent data for independent research, datasets compiled under the old rules are still available in the Databank of the Centre for Economic and Regional Studies (CERS) (https://adatbank.krtk.mta.hu/en/).[4] The mediating role played by the Databank (requesting the linking of administrative data and pre-processing linked data) between NAVÜ and researchers is not necessarily affected by the new law; they can also certainly maintain their existing databases and continue to share them with researchers.[5]

Details about the datasets available in the Databank are published at their website (in English: https://adatbank.krtk.mta.hu/en/adatbazisok/elerheto-adatbazisok/). The Databank maintains two safe rooms and provides remote access via

[3] According to their website, the aim of NAVÜ Ltd. is to improve the international image of Hungary's IT development and to improve Hungary's ranking in the Digital Economy and Society Index (DESI), and the Open Data Maturity Report (ODMR). Hungary ranked 24th in the "Open data" indicator of the "Digital Public Services" dimension in the 2022 DESI.

[4] Note that the new legislation does not impact on the accessibility of datasets already compiled by the Databank, only on the creation of new datasets.

[5] What is not yet clear is whether or not they can continue to make new requests for linking, or NAVÜ would keep this function to themselves. The new legislation is not clear in this respect, so it is expected that this will be decided by NAVÜ's management or by their superiors as they develop their protocols.

a secure server. Researchers cannot bring data from outside and ask for linking, except for data pertaining to territorial units, which are not covered by personal data protection legislation, because the databases are all anonymised and the Databank has no authority to manage linking.

The latest, currently available linked administrative dataset, "Admin3," spans from 2003 to 2017 and includes data from five public registers (healthcare, pensions, taxation, unemployment and public education). The dataset includes monthly information for about 50% of the Hungarian population (children as well as adults) born prior to 2004. A forthcoming linked dataset, to be called "Admin4," will cover 2003–2021 and is expected to be released in 2025.

In general terms, the Databank grants access to their data to CERS staff, their co-authors and students and to researchers affiliated to research institutions accredited by the National Research Fund (OTKA/NKFI) of Hungary, to be used for research and scientific purposes, with differential rules across datasets.[6] Data requests must be made to the Databank using a standard form[7] sent by email. Access is granted on an individual basis (i.e., to an individual researcher and not to the institution to which they are affiliated), for a particular research project.

Access for researchers depends mainly upon the dataset's sensitivity and recency. For example, Admin 3 (covering the years 2003–2017) was accessible to researchers of CERS, their co-authors and students until February 1, 2022, and was made available to others affiliated with an accredited research institution in Hungary thereafter. When released, Admin4 will be accessible to CERS staff only and will be used in projects based in CERS exclusively.

Requests for access are evaluated by considering the applicant's academic record (i.e., their expertise in the topic outlined in the research plan) and their ethical record in using similar data.[8] The Databank may also either refuse or delay granting a request if the research plan overlaps with an ongoing project led by CERS staff.

4 Recent Examples of Using Administrative Data for Policy Analysis

The linked datasets of CERS have been widely used for economic research and have facilitated policy analysis, though mainly outside the government. Several recent projects have exploited the large sample size and accuracy of administrative data, which enables the analysis of small, specific groups or a representative sample of groups that are difficult to reach with the usual sampling procedures and surveys.

[6] See the "Rules of Procedure" published in 2024 https://adatbank.krtk.mta.hu/wp-content/uploads/2024/04/SZ_KRTK_Adatbank_szabalyzat_2024_eng.pdf

[7] https://adatbank.krtk.mta.hu/wp-content/uploads/2024/03/data_request_eng_2024.docx See also https://adatbank.krtk.mta.hu/en/adatvedelem/

[8] Cf. the data access rules of Admin3 https://adatbank. Krtk.mta.hu/wp-content/uploads/2020/02/Admin3_nyilatkozat_adatvedelmi_garanciak_20200922-EN.pdf

Szabó-Morvai and Lengyel (2021) provide a more detailed review of recent research using administrative data in Hungary. It is only possible to include a few examples to illustrate the richness of the available data here.

Researchers from the Finance Ministry have estimated the income elasticity to social security contributions by exploiting the introduction of a family tax credit in 2011, which created substantial variation in marginal tax rates across families according to the number of children (Varga et al., 2020). They did this by using a panel of monthly tax reports that were extracted from the Tax authority's administrative database. Using administrative data enabled them to have a highly accurate measure of reported incomes for both earners in a family, observed just prior to the policy change and three years thereafter, on a sample of over 400 individuals.

Csillag's (2019) research focused on the behaviour of a specific group of workers claiming sick pay: he investigated the extent to which a reduction in the amount of sick pay (relative to wages), introduced in 2011, provided an incentive to shorten the period of sick pay (i.e., to return to work as soon as possible). He used the linked administrative dataset of CERS ("Admin 2") that included, among others, information about employment, sick pay and health insurance events. This estimate would have been impossible to make by using traditional population survey data, given that the treatment group numbered just over 6000 people.

Krekó et al. (2023) exploited a feature of the latest CERS administrative dataset ("Admin 3"); namely, that it also includes information on cognitive skills measured at around age 15 by the National Education Authority, in addition to administrative data about registered unemployment, healthcare and social security records. They estimated the impact of a hiring subsidy designed for young jobseekers aged below 25, introduced in Hungary in 2015, by using propensity score matching with exceptionally rich controls. They showed that the subsidy works well as a screening device by using the information on skills: the programme had the highest impact on youths who had very low levels of schooling, but who scored above their peers on cognitive skills tests.

5 Using Administrative Data in Policy Experiments

There are a few recent initiatives in which administrative data was used in policy experiments that were implemented by the Research unit of the Finance Ministry. The Ministry have access to anonymised tax records and are also in a good position to engage the Tax authority, given that they are controlled by the Ministry. The Ministry's Research unit conducted several experiments focusing on the potential impact of changing the ways or contents of informing taxpayers (Horváth et al., 2020). In one experiment, a random sample of small businesses (using the so-called KATA scheme, which was a form of simplified taxation) was informed about the advantages of paying a higher tax rate (i.e., that they could expect a higher pension if they chose it). The message was varied across the sample and was used to infer the motivation to switch to the higher rate. They found very limited responses, but

there was significant variation across message contents: the share of switchers was highest among those who received a personalised letter and where the prospectus emphasised the potential pension loss, rather than the potential gain.

In another of the Ministry's experiments, a randomly selected sample of small firms received a general warning letter about the prohibition of sales without receipts, the monitoring activities of the NAV and about the risks of being caught. The analysis of value-added tax reports showed a 4% increase in the number of receipts in the month following the warning. This estimate was based on the analysis of thousands of entrepreneurs and tens of millions of transactions, which would not have been possible to collect by surveying entrepreneurs.

6 Barriers to Using Administrative Data for Policy Analysis

The use of administrative data for policy analysis may not gain momentum in Hungary, despite the recent strengthening of the legal and institutional framework, due to the lack of demand for such analysis, the limited capacity of analysts and the infrastructure framework and procedural rules that facilitate access to data.

The basic technical conditions for the exploitation of administrative data for public policy purposes are reasonably good in Hungary, as described previously: the legal framework enables secondary use and the data are of a relatively good quality. The obstacles to better exploitation stem instead from the specific functioning of the Hungarian government. Decision makers and public agencies implementing government policy have little motivation to assess the social and economic impact of interventions either ex ante or ex post. Strong mechanisms to link the social returns of measures to the performance of decision makers are not provided either in budget planning or in other channels of accountability. The expert capacity to carry out analyses, while available in some ministries and in some specialised institutions, is far below what is required for systematic impact assessments, partly because decisions are often not based on the technical arguments of the specialised apparatus. While independent research may continue to flourish (assuming that the new laws are implemented in a favourable way), its focus is determined by academic interests and these interests do not necessarily overlap with the needs of policy making.

References

Büki, Bettina, Ferencz, Bálint, Nyikos, Györgyi. (2022). Közadathasznosítás az Európai Unióban és Magyarországon. In A. Szabó-Morvai, B. Lengyel. (Eds.) *Munkaerőpiaci tükör, 2021: adminisztratív adatok a gyakorlatban, Budapest, Magyarország: Közgazdaság- és Regionális Tudományi Kutatóközpont, ELKH* (pp. 39–46).

Csillag, M. (2019). The incentive effects of sickness absence compensation—Analysis of a natural experiment in Eastern Europe. *Health and Labor Markets* (*Research in Labor Economics, Vol. 47*), emerald publishing limited, Bingley (pp. 195-225). doi:https://doi.org/10.1108/S0147-912120190000047007.

DESI. 2022. *Digital Economy and Society Index 2022 Hungary, European Commission.* https://ec.europa.eu/newsroom/dae/redirection/document/88704

Horváth, K., Keresztély, T., Svraka, A., Tóth, P., & Ván, B. (2020). *Viselkedéstudományi projektek az adóigazgatásban 2019 [Behavioural science projects in tax administration]*. Ministry of Finance. https://ngmszakmaiteruletek.kormany.hu/download/b/52/a2000/Viselked%C3%A9studom%C3%A1nyi%20projektek%20az%20ad%C3%B3igazgat%C3%A1sban%20-%202019.pdf

ITM-BM. 2020. Nemzeti Digitalizációs Stratégia 2021–2030, Budapest, 2020. június. (Partnerségi konzultációra bocsátott, nem végleges változat) Innovációs és Technológiai Minisztérium, Belügyminisztérium. https://2015-2019.kormany.hu/download/f/58/d1000/NDS.pdf

Krekó, J., Munkácsy, B., Csillag, M., & Scharle, Á. (2023). A job trial subsidy for youth: Cheap labour or a screening device? *Social Policy & Administration, 57*(7), 1089–1116. https://doi.org/10.1111/spol.12924

NAIH (2023) Annual report of the Hungarian National Authority for data protection and freedom of information, https://www.naih.hu/files/NAIH_annual_report_2023.pdf

Németh, Ádám (2021) Az adatok tudományos kutatási célú megismerésének jogi szabályozása Magyarországon. In A. Szabó-Morvai, B. Lengyel, B (Eds.) *Munkaerőpiaci tükör, 2021: Adminisztratív adatok a gyakorlatban* (pp. 60–66). Közgazdaság- és Regionális Tudományi Kutatóközpont, ELKH.

Sántha, Gy., & Sikolya, Zs. (2017). Az adatpolitikai fehér könyv társadalmi egyeztetésének tanulságai. *Új magyar közigazgatás, 10.évf*, 1. sz. 46–52.o. https://kozszov.org.hu/dokumentumok/UMK_2017/1/07_Az_adatpolitikai_feher_konyv.pdf

Scharle, Á. (2019). Hungary: A Case Study on Improving Access to Administrative Data in a Low-Trust Environment. In N. Crato, P. Paruolo (Eds.), *Data-Driven Policy Impact Evaluation.* Springer, Cham. https://doi.org/10.1007/978-3-319-78461-8_8

Szabó-Morvai, Á, Lengyel, B (2022) Munkaerőpiaci tükör, 2021: adminisztratív adatok a gyakorlatban (p. 309). Közgazdaság- és Regionális Tudományi Kutatóközpont, ELKH

Varga, G., Nobilis, B., & Svraka, A. (2020). Adóköteles jövedelmek rugalmassága–egy identifikációs kísérlet a családi adókedvezmény 2011-es bevezetése alapján. *Statisztikai Szemle, 98*(1).

Ágota Scharle is a labour economist with over 20 years of experience in quantitative analysis and consultancy in social and employment policy across Europe. Presently she is a senior researcher at the Budapest Institute for Policy Analysis. Prior to that she served as head of research in the Hungarian Finance Ministry, where she coordinated the preparation of a law to enable the use of anonymised administrative microdata for research and policy evaluation.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 15
Building and Accessing a Work and Welfare Longitudinal Dataset from Irish Administrative Data



Hugh Cronin

Abstract This chapter describes the Work and Welfare Longitudinal Database (WWLD), an analytical database in Ireland built from administrative data. The main areas covered are (1) social protection benefits and payments; (2) demographics and relationships; (3) employment and earnings and (4) employment services and programmes. The description of the database's component parts points to its potential for research and policy evaluation. The data available are rich in content, are applicable to analyses of a range of pressing policy concerns and, for many of the key variables, are available over long time horizons. This chapter also outlines some principles for both data access and for maximising the reach of research findings that draw on the database. The WWLD aims to be a fully open-access project, from its source code to any related scientific publications. The chapter concludes by acknowledging current limitations with respect to education and training but, as an ongoing project, successive improvements by harnessing other administrative data sources are anticipated.

1 Introduction

The extent to which EU member states make use of administrative data to ground policymaking in evidence varies considerably. This chapter sets out where Ireland is at on this journey with respect to the development of a large administrative database focussed on both employment and social protection.

The first section sketches the scope and functions of the Department of Social Protection, giving an indication, for an international audience, of the range of

The views expressed in this chapter are those of the author alone and do not necessarily reflect those of the Minister for Social Protection.

H. Cronin (✉)
Department of Social Protection, Irish Government Economic and Evaluation Service (IGEES), Dublin, Ireland
e-mail: hugh.cronin@welfare.ie

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_15

supports and services within its remit. The second section summarises how analytical databases fulfil an organisational need for evidence—no different, perhaps, to many member states, but one grounded in Ireland's experience. This is followed by an outline of the domains that make up the new database and an overview of the principles governing data access and the dissemination of research output. The chapter concludes with some considerations for the future extension and development of the analytical database.

2 Administrative Data and Research in Ireland

This chapter's focus is on the analytical database within the Department of Social Protection (DSP). It is not, of course, the only example of such an administrative data source under the control of a public entity in Ireland. Discussion of several other sources would enhance this chapter but, for reasons of brevity, the interested reader must make do with the following references and links to further information.

Other administrative data holders have developed datasets that facilitate research reports based on administrative data (Revenue Commissioners, 2018) and, in the case of the Higher Education Authority, facilitate access to its administrative datasets for research purposes, subject to certain criteria (HEA, 2024).

At the heart of administrative data in Ireland is the Central Statistics Office (CSO), which takes administrative datasets and joins them. Several State organisations publish statistics and research based on administrative data and some of them share data with the CSO, thereby facilitating the joining of data from disparate sources (Central Statistics Office, 2023). The approach of the CSO is underpinned by the National Statistics Board, which has identified enhancing administrative data for statistical use as one of its priorities for official statistics over the period 2021–2026 (National Statistics Board, 2021). The CSO operates an online system for managing the application, approval and access to research microdata files (both survey-based and administrative).

3 The Operational and Policy Scope of DSP

The DSP administers over 90 separate schemes and services, ranging from social insurance and social assistance income supports, such as pensions, jobseeker payments, illness benefit, disability payments, carer's payments and a minimum income payment, in addition to the universal payment of child benefit, to the delivery of integrated employment and benefit payment services to improve people's job prospects, the operation of Ireland's Social Insurance Fund, a range of employer services including recruitment services, online publication of job vacancies, employment supports (e.g., wage subsidy schemes available to people with disabilities)

and redundancy and insolvency payments. It also manages community services to promote social inclusion.

The registration of births, deaths and marriages also falls under the auspices of the Department, as does the allocation of Ireland's Personal Public Services Number (PPSN)—a unique identifier used to access social services. Figure 15.1 illustrates the organisational structure, according to the European Commission's Mutual Information System on Social Protection (MISSOC).

With such a breadth of schemes, policy areas and services, and with such vast coverage, encompassing a large majority of people in Ireland at any given time, a wealth of data are generated by DSP's activities. The corollary is that the demand for analytical output is considerable. The DSP needs to have point-in-time reporting on

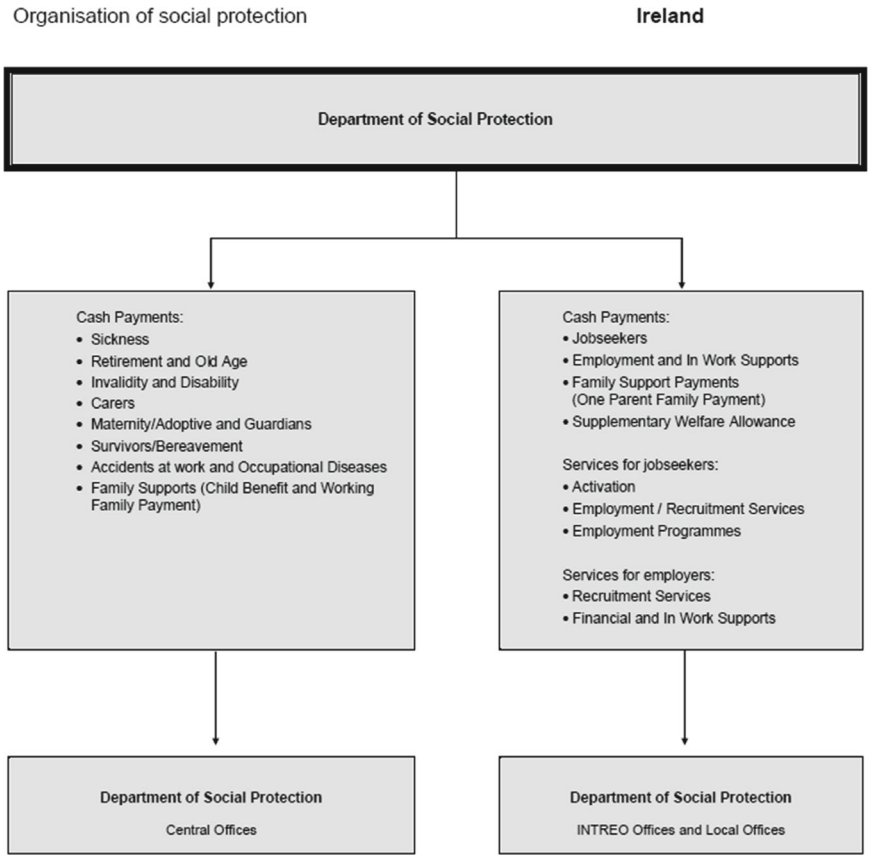


Fig. 15.1 MISSOC's graphic on the organisation of social protection in Ireland. *Source* Mutual Information System on Social Protection. *Note* Intreo is the "one-stop-shop" integrating income maintenance and employment supports, encompassing Case Officers who are directly employed by the Department of Social Protection and the external contracted service providers delivering services in over 150 locations

current caseloads, processing times and claim volumes. It also requires the capacity to examine how people in receipt of its income support and employment services fare over time.

As one example, with Active Labour Market Policies (ALMPs) playing a critical role in reconnecting people to the labour market, it is essential that high-quality analysis underpins referral decisions and caseworker guidance. At the level of individual programmes, a greater understanding of programme impacts will help inform the menu of interventions that should be on offer, the scope of eligibility and the cohorts that caseworkers should direct towards particular programmes. The points in the economic cycle at which particular programmes have specific benefits could also provide useful input to these referral decisions.

At a broader level, the Public Employment Service (PES) also needs evidence about how to structure sequences where it needs to take a longer-term approach to co-ordinating a series of services and programmes, helping people to add new skills over time and building on previous programmes or qualifications. With access to large-scale administrative data on unemployment, a series of PES interventions could be based on the employment trajectories of people with similar characteristics.

There is a critical demand for evidence about the net impact of services and programmes on people's long-term employment prospects as reflected in earnings, hours of work (when available), weeks of work, reliance on income support and various combinations of labour market statuses over time, where these may reflect improving or disimproving trajectories. The analysis requires the appropriate application of relevant techniques and a solid interpretation about how the data relate to the reality that they represent. However, a foundational aspect is high-quality longitudinal data, where relevant background characteristics can be controlled for and a suitable suite of long-term outcomes can be measured. It is essential to be able to identify a set of individuals whose characteristics unite them, in some way or to some degree, at a point in time and to trace their outcomes at repeated points in the future.

While official statistics and survey-based research datasets remain a critical part of understanding the policy domain in which the DSP operates, the Department's administrative data also have great potential for facilitating the longitudinal analysis needed to inform policy choices and to illustrate the outcome of its practices.[1]

With this in mind, and cognisant of the fact that any use of administrative data for analytical or research purposes is a by-product (as administrative data almost always is) of the essential role played by the DSP, the Department embarked on a project to rebuild its longitudinal analytical database in 2022.

[1] Some point-in-time outputs from DSP administrative data have already been designated as national statistics. See the DSP Quarterly Statistics Report (Department of Social Protection, 2024).

4 Learning to Build Analytical Databases

The Work and Welfare Longitudinal Database (WWLD) sits within a new centralised data repository and modern analytics platform, with the ambition of ensuring that the data refresh of the WWLD is automated, that the access protocols are trusted and that the answers provided to analytical questions are replicable.

The ambition of the WWLD builds on lessons learned from a previous analytical dataset, the Jobseeker Longitudinal Dataset (JLD), which was first produced in 2012 using a legacy technology (SAS). This also merged claim information, data from the training provider at the time (FÁS, now renamed SOLAS), employment data and demographic information, with a monthly refresh cycle. As the DSP's data infrastructure developed, management of some schemes moved from legacy platforms to newer platforms but without a corresponding update to the JLD pipeline. Inevitably, this led to various elements of the dataset becoming outdated.

Analysts had to engage in bespoke data compilation and extraction for each analytical project as the historical dataset decayed. While ad hoc solutions were a temporary fix, the level of effort and the scope for variation in different approaches brought the need for a replacement dataset into focus. The initial steps towards a new dataset were flagged in Pathways to Work 2021–2025, the national employment services strategy, with Commitment 83 pointing to the need to "Conduct a data audit exercise on a subset of DSP administrative data to assess their comprehensiveness, consistency, reliability and range with a view to facilitating longitudinal analysis of outcomes."

In parallel, the DSP's data analytics capability and data management practices matured, with a greater focus on open-source tools as analysts developed capacity in Python, R, and SQL. The development of a new research database presented an opportunity to implement modern approaches as a means to future-proof the (potentially valuable) resources represented by the DSP's data (Fig. 15.2).

There was an opportunity to improve data quality based on the experience of the Department's analysts in content as well as in form. The JLD had several limitations, despite its undoubted reliability. There was a greater sense of ambition in designing the WWLD and this was abetted by a larger number of analysts familiar with DSP data and the analytical ends to which it could be put. This was enhanced by external consultation with previous users and policy areas with new policy challenges that needed analysis to inform choices.

The result is a multifaceted domain-led database, where a variety of sources have been carefully collated and joined. It covers large swathes of DSP's remit, from registered births through the allocation of Ireland's PPSN and proxy measures of arrival in the country (the allocation date of that number and recorded alongside the country of birth) for 9.9 million unique identifiers. This gives a reasonable approximation of the population at a specific point in time (although there are circumstances in which people may not need to register for the PPSN).

The WWLD's careful construction was the result of the painstaking work required to collate a range of disparate data sources from separate operational sources. The effort required notwithstanding, it is worth noting the fortunate starting position for

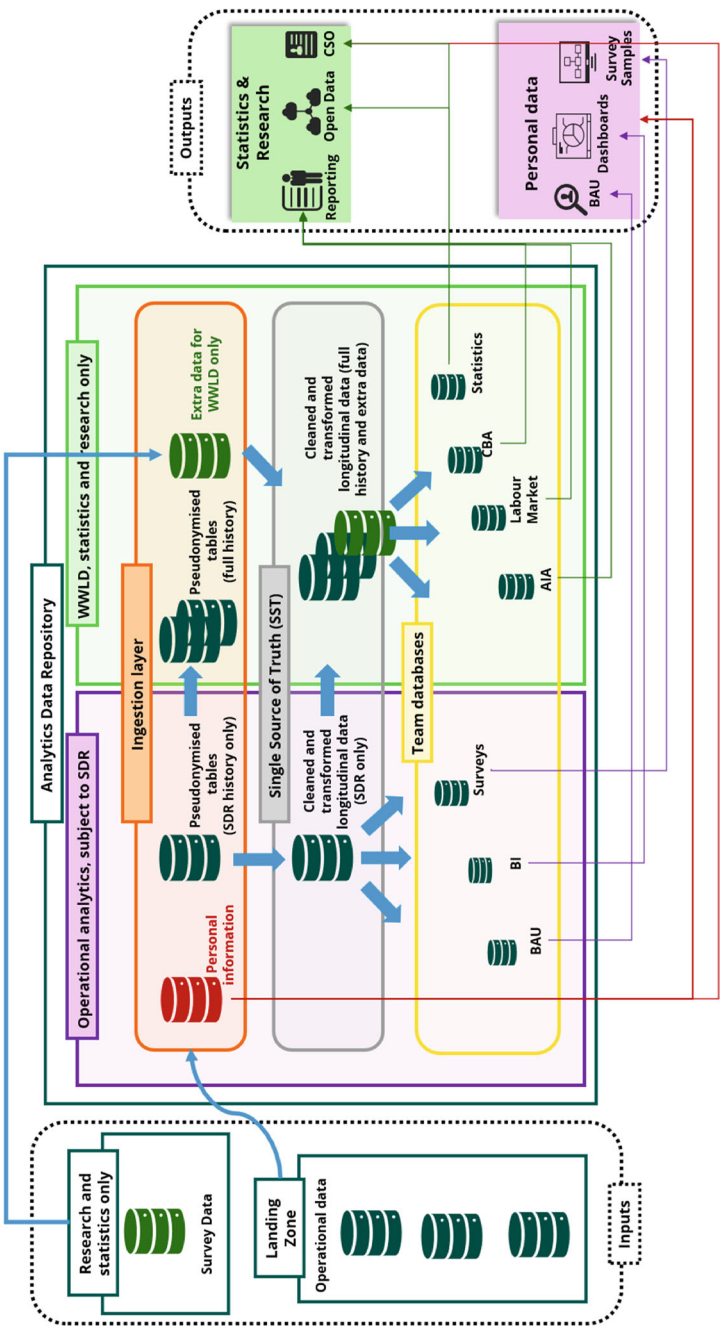


Fig. 15.2 Data schematic. Legend—SDR: structured data retention; CSO: Central Statistics Office; BAU: Business Analytics Unit; BI: Business Intelligence; CBA: Cost Benefit Analysis; AIA: Actuarial Investment Analysis; ADR-SST: Analytics Data Repository Single Source of Truth

the construction of this type of database. The institutional structure of DSP's remit is at the national level, meaning that data are collected nationally rather than across sub-national levels. This differs from the structure of many EU member states.

The DSP's responsibility for social insurance and its records of employment contributions, which are validated as people make claims for insurance-based pension entitlements, means it has a rich record of employment in Ireland. There are paid social insurance contributions dating as far back as 1932, unemployment benefits information from the 1970s and earnings data from the 1990s, as well as detailed information about the supports and services that comprise Ireland's social protection system. The following section looks at the collation of the data sources in greater detail.

5 The WWLD's Domains

One of the WWLD's key design challenges involved maximising the utility of the information collected and ensuring that different domains and data sources, thereby allowing for adjudication between potentially conflicting data points. As reflected in DSP datasets, people's lived experience over time can be extraordinarily complex, far more so than is suggested by surveys or point-in-time snapshots. The project design team's approach was to consider a range of data sources, to make informed decisions based on eligibility criteria, administrative rules, operational practices and data quality and to then codify a shared best estimate of how to interpret the information. The end product is necessarily imperfect, but the best estimate from a range of clues as to what transpired.

Notwithstanding the range of data sources available, even a broad range of data on employment and income support does not provide a limitless view of the population, nor a definitive account of each individual's situation at a point in time. Absences from administrative data could reflect either, for example, a person who emigrated to another country or a person who became economically inactive. Moreover, adjudicating between these competing explanations for the absence can be challenging. Even for those who remain in the country over a period of time, employment can overlap with a range of other statuses (caring responsibilities, part-time unemployment support and disability).

As an example, the DSP's work on employment data illustrates the investment required in developing a reliable database for analytical purposes. Prior to 2019, Ireland's taxation collection authorities, the Revenue Commissioners, produced an end-of-year file covering all taxable earnings within a given year. However, while the data may be sufficient for establishing a tax liability, the quality of employment start and end dates is unsatisfactory. In some cases, simple duplicates are easily corrected; however, more ambiguous scenarios are more complex to address, such as where one employment period is contained within a larger employment period with the same employer, or where two periods with the same employer each have a missing start date or end date.

The PRSI data, which is collected by the Revenue Commissioners on behalf of DSP, does not have a sub-annual data component—typically, the relevant information is the social insurance class and the number of weeks of contributions in a calendar year. While these can be interpreted without the employment start or end dates, they can also provide a clue as to the duration of employment. Contributions data are leveraged for the WWLD in order to probabilistically allocate periods of employment within a calendar year or across contiguous years.

Since 2019, employers have been required to report employees' pay and deductions in real-time—each time they pay their employees—to the Revenue Commissioners. This means that administrative earnings data on earnings are now available at a periodicity that is more in line with administrative data on unemployment benefit claims. It also means that the payslip data can be used to generate more reliable start and end dates for employments ongoing in 2019.

The payslip data could reduce reliance on the end-of-year file over time. In the employment domain, imputation draws on payslip and contributions data to improve employment start and end dates. This is critical to both identifying point-in-time status and for employment-based outcomes. For an example of the application of imputed status at points-in-time, see the chapter "Sequences of support schemes in Ireland" in (OECD/Department of Social Protection, Ireland/European Commission, Joint Research Centre, 2024).

A further refinement would be to layer social welfare claims (measured to the precise day) in order to further refine our understanding of employment periods. This provides information about the particular contingency that the person experienced, the occupational background and household composition as well as important information about the frequency and duration of previous episodes of unemployment.

Moving beyond the fact of receiving a particular type of income support, the next layer of information is the employment services provided to the person, as well as any referrals to ALMPs. How and when this occurred may be relevant to showing how it helped the person's employment prospects. A further avenue of inquiry is the examination of what kinds of firms assist the exit from unemployment to employment, through the WWLD's linked employer-employee facet. A previous source for this kind of research—a survey-based, linked employer-employee dataset produced by Ireland's national statistics institute (the CSO)—has been discontinued.

All of this is complemented by a layer of data that is more akin to a survey dataset, albeit a survey for which the target is the population, rather than a sample, of newly registered unemployed people. The Probability of Exit (PEX) questionnaire asks a series of socio-economic questions about a jobseeker's background and skills that provide useful contextual information for analytical purposes, some of which point towards the archetypal unobservable characteristics where administrative data are, typically, weaker (such as whether jobseekers are willing to move location to take up a job).

By investing effort in raw data sources, and by collating several separate indicators of a person's labour market status, there is the real possibility of having a very rich, nuanced and coherent view of people's (often complex) trajectories, while also being able to report succinctly on aggregated labour market statuses at a point in

time. This enables a range of analyses to be conducted, thereby facilitating a greater understanding of both the social protection system and the labour market.

One final facet worth noting concerns the attempt to move beyond the basic administrative view of the claimant as a sole entity and towards a household view, by taking a longitudinal approach to household composition. To this end, the WWLD has domains dedicated to dependants and relationships. Typically, administrative social protection data are structured around the recipient. The WWLD adds a layer of nuance and richness when trying to represent, for example, labour supply decisions that are, at least to some extent, taken at a household level by creating two-way links based on relationships and the additional benefit payments made in respect of dependants. Similarly, time spent as a child dependant is not something captured in the source data but can be recreated through a similar set of links. This provides an opportunity for longitudinal perspectives on the components of household income that relate to child poverty outcomes (Table 15.1).

5.1 Making the Most of Administrative Data: Principles of Access

The WWLD will be an invaluable asset that will provide a single source of truth about the information that the DSP has by generating an administrative data perspective on concepts that have previously only existed in sample-based surveys.

The WWLD's design and delivery has been a project many years in gestation and the objective now is to ensure that the use of the data asset maximises its power. To this end, the DSP will continue to make its data available for relevant policy questions. A database as powerful as this is undoubtedly attractive to a large number of users. Good quality administrative data allow for the application of sophisticated techniques that can be applied to a range of sub-groups. Researchers can apply cutting-edge methodologies, with the vast range of variables providing ample capacity for addressing problems of confounding that might otherwise stymy efforts at causal interpretation.

The DSP needs to, in the first instance, maintain its position as the custodian of data when supporting this analytical activity, through the building of the WWLD.

This is reflected in two aspects –

(i) The steps that the DSP has taken to assimilate raw operational datasets into an analytical dataset and, in doing so, to minimise the granularity of data and to abstract from the personal data collected as part of social welfare claims
(ii) The access that DSP grants to this database and its oversight of users.

While (i) is of critical reputational importance to the DSP, it is detailed in the data protection impact assessment document produced in advance of the commencement of work on the WWLD, which outlines the data pseudonymisation and obfuscation steps. For reasons of space, the focus in this chapter is on (ii), specifically, and on the principles that inform access to the WWLD.

Table 15.1 Outlines the domains under development and compares them to what was available in previous research datasets

Domain names	Variables include	Coverage in the JLD	The improvement the WWLD makes on the JLD
Demographics	Obfuscated date of birth, sex, marital status, nationality, country of birth and county of residence	Complete	Additional data points, improved pseudonymisation, proxies for migration flows
Employment	Employment start and end dates, employer identifier, NACE sector, number of employees, gross earnings, social insurance (PRSI) class and weeks of insurable employment, credited social insurance contributions	Partial	Greater accuracy of employment start and end dates through imputation
Payslips	Employment start and end dates, employer number, PRSI class and weeks of insurable employment, gross pay per payslip, and pay frequency from employer	None	
Relationships	Relationship type, relationship start and end date, related customer in this relationship based on registration of births, deaths, and marriages and income support increases in respect of dependant adults	Partial	Additional data points, longitudinal view
Claims and Entitlements	Start and end dates of entitlement and a claim, total weekly amount, payment item types, add-ons (various allowances such as the fuel allowance, over 80 s allowance, island allowance, living alone allowance)	Complete	More claim types, transformations that better represent reality
Dependants	Dependant type, start and end dates, link to or from dependant or claimant, link to claim	Partial	Longitudinal view of the two-way relationship between the claimant and dependants and, for the first time, a household perspective by linking adult and child dependants and providing multigenerational links

(continued)

Table 15.1 (continued)

Domain names	Variables include	Coverage in the JLD	The improvement the WWLD makes on the JLD
Probability of Exit (PEX)	Probability of Exit score assigned to each new claim, date assessed, answers to all PEX questions (i.e., components of the score)	None	Incorporation of a set of questions closer in nature to a survey dataset, albeit a survey where the target is the population, rather than a sample, of newly registered unemployed people
Payments	Payee type, scheme code, scheme description, payment item, payment item amount, total payment amount, payment dates, payment method, payment status, dates of arrears	Partial	Higher frequency, greater detail
Means	Data on the means testing processing underpinning the social assistance system, including means categories, dates of means assessment, date of effect, associated claim IDs, and the net weekly means amount once disregards have been applied and the claimant's (and sometimes the spouse's) income and capital have been assessed	None	
Statistical codes for health data	Aggregated values based on the International Statistical Classification of Diseases and Related Health Problems (ICD-10)	None	
Employment services and activation	Dates of referral to employment services, referral to external employment services providers	None	
Imputed Status (e.g., ILO status)	A derived field estimating, for example, the International Labour Organization status or the Live Register status that corresponds most closely to the demographic, employment and claim information available for this time[2]	None	

The DSP's first responsibility is to ensure that it grants access only to individuals or organisations with a capacity for secure data storage, an understanding of disclosive analytical outputs and a track record in high-quality data analysis. The DSP will aim to model access requirements on those used by the national statistics office.

[2] See the CSO description of the Live Register, and the ILO explanation of ILO status.

In setting out how the access arrangements will work, it is worthwhile restating some of the facts of the WWLD project. Substantial public funds have been directed to the collection of the data and to the transformation from multiple sources of raw, operational data into a curated analytical dataset. The objective of such an asset is to better facilitate shared understanding of the people whose reality these data (partially) reflect: their situations, their choices, the State supports they avail of, their outcomes and, perhaps most importantly, the causal interaction of these aspects. Given this origin and objective, the data, code and all outputs from its analysis should be open, transparent and reproducible.

The rationale for such an approach builds on the characterisation of microdata as a public good (a non-excludable and non-rivalrous good) as described in Crato and Paruolo (2018). The model of working with the database for research purposes is essentially collaborative—collaboration with the people who build the database and collaboration with those whose analysis either precedes or follows a given research question.

There are three levels at which such an approach could yield results. Firstly, in its construction phase, the WWLD has been designed using open-code repositories. This means that the design choices that generate the database are open to scrutiny and any analyst with in-depth knowledge of the domain can examine the design structure and can propose improvements to the production code. The care and consideration given to the WWLD's design, and the way in which it reflects the DSP payments and services, does not mean that a better design is impossible; rather, facilitating an open-code approach to its construction reflects the designers' openness to surfacing any improvements that can be made.

Second, beyond the database design, at an analytical level, open code repositories mean that analysts can generate code as they work and that this code is available for review and critical evaluation. This allows a greater degree of transparency—critical for data protection purposes—and means that the analyst can move more rapidly to the later stages of analysis by minimising unnecessary duplication across research projects.

Third, a metadata management platform has been incorporated into the WWLD project and will be available to analysts using the data. This documents the data lineage of the attributes in the WWLD, while also providing critical business understanding of the database attributes. As analysts work in specialised areas of DSP's remit, they can add their observations and understandings, thereby avoiding knowledge silos and ensuring critical business knowledge is not lost as people leave the DSP.

As each researcher progresses, she benefits from the development of previous code and metadata by others. Again, this does not presuppose an unthinking following of previous analyses; any superior approach to modelling a particular outcome or highlighting an association between variables builds upon the acknowledged efforts of predecessors. Making the code developed by previous analysts available enshrines reproducibility as one of the central tenets of the research facilitated by the WWLD.

Any researcher who is used to guarding access to data as a competitive advantage and treating code as intellectual property may consider that this open code approach

grants an unusual, and perhaps unfair, advantage to later users. However, acknowledging previous analytical work, through citation, will also be a prerequisite of access to the database. Each researcher benefits from, and acknowledges, the efforts of previous investigations. In short, the ambition is to marry the in-depth domain knowledge and empirical rigour of investigative research work with the transparent practices of open data software developers.

This approach is grounded in the fact that the benefits of development investment by the DSP will be shared with researchers and they, in turn, are expected to share the benefits of their empirical investigations. Undoubtedly, this method increases the transparency of the analytical endeavour and, across the breadth of data users, improves overall quality. Any questionable selection criteria or modelling choice can be queried and robustly defended on an empirical basis, provided that all code is available; competing explanations can be contrasted and the merits of each can be fully presented. This can only be to the benefit of rigorous, well-considered research.

As with database design, where all ideas for better data design and construction can be tested, all modelling strategies at the analytical phase can likewise be tested. Sharing code, in this context, is a strong statement of confidence and integrity, where an advance across the whole research field is the valuable objective that is being pursued by all contributors. The EUPL, the European Union's open-source software licence, or any of those on its compatibility list, will apply to all code generated by researchers using the WWLD.[3]

While this describes the principled approach to access, more practical details can also be sketched out. The first step involves a detailed proposal of the research questions without a commercial application. This must include an initial attempt, using available metadata, to identify the relevant population and must include an outline of how this population relates (as narrowly as possible) to the question being investigated. Once agreed, remote access to the analysis-level data will be granted provided that the relevant researcher data access and data confidentiality agreements have been signed. It is not envisaged that a fee will be charged for research access.

5.2 Making the Most of Administrative Data: Transparency from Raw Data to Research Findings

The corollary of this open approach to both how the database is constructed and how the analysis is carried out is that the research outputs should also be widely available. The development investment by the DSP is based on data that cover large swathes of the population. If policy improvements can be suggested, on the basis of the analysis of these data, then the results should be widely available for consideration, rather than sequestered in pay-walled journals. Good quality analytical outputs will be of far greater concern than a restriction to a sliver of the population can justify. The

[3] A list of compatible licences is provided in Article 5 EUPL. See COMMISSION IMPLEMENTING DECISION (EU) 2017/863 in the Official Journal of the European Union.

policy discussion that such analyses will inform affects a variety of people, their employment prospects, household income and child and family outcomes.

All output from external users of the WWLD will be published as open-access articles to reflect the belief that the research output is something that should be widely available. As part of data access agreements, it will not be possible for researchers to access the WWLD database—itself a reflection of substantial public investment and collaborative practices—and subsequently to publish findings as pay-walled articles in subscription-based journals. This is in keeping with the UNESCO Recommendation on Open Science and with Open Research Europe, which provides an opportunity for researchers, funded through a variety of EU sources, to publish their latest findings at no cost to themselves.

At a national level, this approach can be traced back to the requirement for Irish scholarly publications that arise from publicly funded research to be openly available from 2020, and the current policy as outlined in Ireland's National Action Plan for Open Research 2022–2030 and Impact 2030, Ireland's research and innovation strategy. It is also reflected in the practices of Research Ireland, the largest funder of competitive research in Ireland, which is a signatory of Plan S and a member of cOAlition S.[4]

These prescriptions for accessing the WWLD are founded on a belief that the highest quality data and the most relevant and robust research outputs are facilitated by an open research approach, where collaboration is encouraged to improve both the database and the encoded methodologies applied to the data within it. In turn, this approach facilitates reproducibility and brings integrity and transparency to the fore in the research process, thereby ensuring that the maximum impact can be realised from every research endeavour.

6 Future Developments

Despite the range of analyses facilitated by this careful curation of data sources, the WWLD is not a finished product. The initial set of domains will provide critical information about people's labour market status over contrasting times—times of high unemployment and times of record high employment—while providing specific details about relevant characteristics to control for confounding factors in causal analysis.

The current development work has focussed primarily on internal datasets—income support, employment services and the demographic and household characteristics. Useful additions to the core dataset are already envisaged—these include administrative datasets on education and training, which are critical factors for social protection and labour market outcomes, but for which only partial coverage

[4] See Taighde Éireann Research Ireland (https://researchireland.ie/), which was established in August 2024, following an amalgamation of Science Foundation Ireland (SFI) and the Irish Research Council (IRC). SFI was the signatory to cOAlition S (https://www.coalition-s.org/).

exists at present. It is also anticipated that the WWLD's component parts may be upgraded over time, such as expected improvements to administrative employment data through the addition of variables capturing the number of hours worked.

There may also be further developments in the ways in which people interact with digitalised information systems for the delivery of public services. Over time, it is hoped that the data quality will continue to mature and that the codebase will continue to grow. In parallel, the principles governing its access and outputs based on its data may be further enunciated and explained in line with overarching legal, administrative and data protection structures.

Most importantly, there will be an ongoing process of collective improvement of the data, of modelling approaches and of the analytical interpretations of the data, all of which will harness the analytical capacity and unique perspectives of researchers motivated to work with such extensive, comprehensive and policy-relevant data.[5]

References

Central Statistics Office. (2023, February 24). *Arrivals from Ukraine in Ireland Series 9.* Retrieved from cso.ie: https://www.cso.ie/en/releasesandpublications/fp/p-aui/arrivalsfromukraineinirelandseries9/contactdetails/

Crato, N., & Paruolo, P. (Eds.). (2018). *Policy, Data-Driven Policy Impact Evaluation—How Access to Microdata is Transforming Policy Design.* Springer Open.

Department of Social Protection. (2024, July 19). *Quarterly Statistics.* Retrieved from gov.ie: https://www.gov.ie/en/collection/b08c7-quarterly-statistics/

HEA. (2024). *Research Microdata Files (RMFs).* Retrieved from Higher Education Authority.

National Statistics Board. (2021). *Strategic Priorities for Official Statistics 2021–2026.* Dublin: Government of Ireland.

OECD/Department of Social Protection, Ireland/European Commission, Joint Research Centre. (2024). *Impact Evaluation of Ireland's Active Labour Market Policies.* OECD Publishing.

Revenue Commissioners. (2018, August 01). *Research reports.* Retrieved from revenue.ie: https://revenue.ie/en/corporate/information-about-revenue/research/research-reports/index.aspx

Hugh Cronin is an economist with the Irish Government Economic and Evaluation Service (IGEES) in the Department of Social Protection. He is based in the Budget and Means Policy Unit, having previously worked in the Labour Market Analytics Unit, which delivers counterfactual impact evaluations under the Government's Pathways to Work strategy. He has worked in the Department since 2015, with the exception of 2021–2022 on secondment to the Department of Finance as part of the secretariat to the Commission on Taxation and Welfare.

[5] Anyone interested in this collaborative approach should contact LMAnalytics@welfare.ie and StatsUnit@welfare.ie.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 16
The Long and Winding Road: The Access to Administrative Data for Research in Italy



Roberto Leombruni and Chiara Ardito

Abstract There is a tradition in Italy regarding the use of administrative data for research purposes that dates back to the 1980s. At present, however, the situation is quite scattered among several heterogeneous and often uncoordinated initiatives, even when the data source is the same. As a consequence, we are far from an optimal exploitation of the potentialities presented by administrative data sources and researchers often face a long and uncertain road when looking for research data for policy evaluation. In this chapter, we provide an overview of the two main options that are currently available to researchers interested in studies on labour and health and, to a lesser extent, education. The first is a setting in which data production and access is centralised by both the Ministry of Welfare and the National Social Security Administration. The second is a wide range of initiatives that are characterised by cooperation between public administrations, universities and research institutions. We conclude by discussing the main issues that characterise the current situation and the directions that may be followed for a better exploitation of administrative data for research.

1 The Long and Winding Road

The use of administrative data for scientific research has a long tradition in Italy, with a series of studies having been published around forty years ago on matters such as firm demography, job creation and destruction and earnings differentials (Contini & Revelli, 1987, 1989). This was possible thanks to *ad-hoc* agreements

Roberto Leombruni acknowledges a European NextGenerationEU funding through the PNRR project “Age-It - Ageing well in an ageing society”, code PE0000015, MUR 341/2022 tender.

R. Leombruni
University of Torino, Torino, Italy
e-mail: roberto.leombruni@unito.it

C. Ardito (✉)
European Commission Joint Research Centre, Ispra, Italy
e-mail: chiara.ardito@ec.europa.eu

© The European Union and The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_16

whereby the National Social Security Administration (INPS) granted access to its data to a pool of researchers based at the University of Turin. At the turn of the century, this undertaking eventually expanded into further collaborations between the INPS and other universities and was focused on the creation, use and dissemination of a 1% linked employer-employee longitudinal sample of Italian workers. This triggered a wave of studies into a wide range of topics from workers mobility and precarity to gender differentials, retirement decisions and the pension system fairness, the evaluation of employment protection regulations and welfare provisions. At the same time, this setting was characterised by easy access to simplified, anonymised samples based on INPS' data, and by the (more difficult) access to more comprehensive versions that required an affiliation to an institution that had undertaken a formal agreement with the INPS.[1]

This setting has changed radically over the last two decades. On the one hand, the wave of regulations on personal data protection has made it more difficult to both link and disseminate microdata. On the other hand, the collaboration between the INPS and the universities to develop research databases has slowed down in favour of their centralised production and dissemination.

The situation that has emerged is twofold. In a centralised setting, primarily for nationwide studies in labour economics and welfare policies evaluation, researchers can access population-level data through research data centres located at INPS premises or through simplified, anonymised samples. INPS samples are produced directly by the INPS on behalf of the Ministry of Welfare.

Beyond the centralised setting, other initiatives have actually experienced a boost, driven by two main factors, one data-driven and the other research-driven. The first stems from the introduction of a new data source, the so-called Mandatory Communications, jointly managed by central and regional institutions. This dataset tracks every start, transformation and termination of employment contracts. Its richness, especially compared with the previously scant availability of microdata representative of local areas, has triggered a wave of studies on the functioning of regional labour markets, often based on close collaboration between regional institutions and universities.

The second factor driving the increased use of administrative data, the research-driven one, can be traced back to the seminal work done by the World Health Organisation's commission on the social determinants of health (Marmot, 2005). After the 2007 economic crisis, which exacerbated social inequalities, research on the social determinants of health has gained substantial political relevance and attention, both internationally and in Italy. This has facilitated the launch of several initiatives linking socio-economic data to various health registers maintained by public administrations.

[1] It is out of the scope here to propose any selection of studies published during that phase. The most diffused among the databases mentioned was the Work Histories Italian Panel (WHIP), developed by the University of Turin in cooperation with INPS and the Laboratorio Riccardo Revelli, which was re-distributed to about one hundred Universities and research centres in a simplified version. Other agreements that centred on INPS data involved the National Institute for Public Policies Analysis (INAPP, formerly ISFOL), the University of Roma La Sapienza, the Bank of Italy and the Bocconi University.

In the sections that follow, we review the data and data accessibility characterising both the centralised setting (Sect. 2) and a selection of collaborative and regional initiatives (Sect. 3), discussing the main challenges of the current system and potential avenues to improve the systematic use of administrative data for research purposes.[2] Our focus is mainly on data and studies concerning working careers and the social determinants of health, and we limit our review to settings in which data access does not entail monetary payment—thus excluding, for example, business repositories that sell extracts from their administrative archives. Nor do we cover the numerous efforts undertaken by ISTAT (the Italian National Statistical Institute) to promote the broader use of administrative data, some of which—such as resident registers or the register of public pension recipients—are available to researchers on-premises through the "ADELE" laboratory.[3]

2 The Centralised Setting

The centralised setting allows for the use of administrative data on employment and welfare in two ways: the first is by means of three national representative samples that are delivered to a requesting research institution directly and the second by means of direct access to population data in the INPS premises.

2.1 Working Career Microdata Delivered to Research Institutions

The so-called Fornero Reform of the labour market (Law n. 92/2012) aimed to ensure the independent evaluation of its effect and assigned to INPS and to the National Institute of Statistics (ISTAT) the responsibility of compiling anonymous microdata on employment, wages and social benefits and to make them available for scientific purposes (Art. 1, § 4–5). The main goal was to standardise and centralise the dissemination of social security data, thereby discontinuing the previous system of case-by-case agreements between INPS and the research community.

[2] It is worth noting that a rich body of high-quality research is also emerging from other individual initiatives and institutional collaborations on topics related to those reviewed here. Examples include studies on labour market heritability using twin data (Brescianini et al., 2024), the effects of immigrant legalization on crime (Pinotti, 2017), the evaluation of health care policies (Bruni et al., 2016) and costs (Berta et al., 2021), and the distribution of personal wealth based on inheritance tax records (Acciari et al., 2024). Together, these contributions underscore the vast potential of broader access to administrative data for research.

[3] See: https://www.istat.it/en/data/microdata/adele-laboratory/. On the current status of ISTAT activities on the matter see, e.g., Di Zio et al. (2022).

The implementation of this norm resulted in the publication of two longitudinal samples: the LoSai (Longitudinal Sample INPS) database,[4] a 7% representative national sample of INPS' data on work biographies from 1985 onwards, and the CiCo (Campione Integrato Comunicazioni Obbligatorie) database,[5] a 14% representative sample of the national "Mandatory Communications" archive, which records employment episodes from 2008 onwards. The structure and content of the "Mandatory Communications" database are described in more detail in paragraph 3.1 of this chapter.

At the time of writing, access to both the *CiCo* sample of the "Mandatory Communications" and the *LoSai* database derived from INPS archives is suspended. Following the discontinuation of *LoSai*, INPS has made available a new dataset for research purposes—the *MFR* (Microfile for Research). Like *LoSai*, it is based on INPS administrative archives and provides linked information on social security, employment, and pension histories. However, the *MFR* differs in several important respects: it covers a shorter time span (2009–2014) and adopts a different unit of analysis, offering yearly summaries at the individual worker level rather than data at the job-contract level. It also provides less granular detail on certain key variables—for instance, age is reported in only three broad categories. On the other hand, it includes information that was previously unavailable in *LoSai*, such as the geographical area of employment.

Access to these datasets is generally relatively straightforward and transparent, but restricted to researchers affiliated with institutions recognized by COMSTAT or EUROSTAT (such as universities, research centres, or international organisations like the JRC or OECD).[6] The application procedure entails submitting a research proposal—either to INPS (for the MFR) or to the Ministry of Welfare (for CiCO and LoSai when they were available). Importantly, applications are evaluated solely on formal eligibility criteria, without any assessment or prioritization based on the scientific relevance or quality of the proposed research.

Despite the relative ease of access enabled by this first centralised policy, its impact, particularly in the case of the LoSai, was limited. The LoSai database was created through a basic extraction from INPS' administrative archives, strictly following the short list of variables mentioned in the law, with minimal effort to adapt the dataset to the purposes and needs of scientific research. This likely reflects the "no budgetary impact" clause included in that same Law (Art. 1, §6). As a result, the information content proved insufficient for many scientific purposes. For example, the geographical dimension was limited to the region of residence, as observed at the end of the twenty-year panel, an important limitation given Italy's high internal

[4] For additional information on the LoSai sample, see, for example, Casarico and Lattanzio (2023) and Berton et al. (2023). To the best of our knowledge, no official webpage is currently available, as the dataset has been discontinued.

[5] For additional information on the Mandatory Communications (CiCo) sample and details on data accessibility, see: http://dati.lavoro.gov.it/microdati-la-ricerca.

[6] For details on the recognition procedure as a research institution, see the relevant COMSTAT directive available through INPS at https://www.inps.it/it/it/dati-e-bilanci/File-per-la-ricerca-MFR.html.

mobility and pronounced regional labour market heterogeneity.[7] Some of the initial pitfalls were subsequently addressed, leading to a modest increase in usage before the database was discontinued. Nevertheless, during its period of availability, LoSai remained significantly underexploited relative to its potential[8]—and its removal from the pool of accessible research data represents a missed opportunity for empirical research on the Italian labour market.

Another valuable data resource for research at the intersection of labour, income, and health is the AD-SILC database, which links administrative records from INPS with the Italian survey component of the European Union Statistics on Income and Living Conditions (EU-SILC) (see, e.g., Conti et al., 2023; Raitano, 2012). The dataset was developed as part of the T-DYMM microsimulation model of the Italian social security system by the Department of the Treasury within the Italian Ministry of Economy and Finance and the Giacomo Brodolini Fundation.[9] To the best of our knowledge, access to the database is currently not publicly available.

2.2 Secure Access in Research Data Centres

Since 2015, the National Institute for Social Security (INPS) has expanded scientific access to its extensive administrative archives through the *VisitINPS Scholars Programme*.[10] The programme provides on-site access—via a secure platform called *VisitINPS*, available in INPS regional offices—to detailed anonymised microdata on workers, firms, employment spells, earnings, contributions, benefits, and pensions. Access is granted through a periodic competitive call for proposals open to scholars affiliated with universities, research institutes, or public institutions. Proposals are evaluated by an internal INPS scientific board based on their scientific relevance and consistency with the Institute's research priorities.

The programme makes available to researchers a wide range of datasets, all of which include anonymous identifiers that allow linkage across archives, including: (i) the pension-benefit archive (1995–present); (ii) an employer–employee panel for private-sector workers (1974 onward, with a monthly component since 2005);

[7] In this respect, an example is provided by Berton et al. (2023). Due to the absence of information on the region of work in the LoSai dataset, they were compelled to merge it probabilistically with the CiCo database in order to analyse regional disparities in the effects of the *Jobs Act* reform. The full merging procedure is documented and available in the paper's online repository.

[8] A quick bibliographic search on Google Scholar using only the keywords "Italy" and "LoSai" returned 38 results in the first six years following the dataset's introduction, increasing to 102 in the subsequent six-year period. A similar search on EconLit identified only 7 indexed papers that made use of the LoSai sample.

[9] More information on the projects that led to the development of the AD-SILC dataset can be found: https://www.dt.mef.gov.it/export/sites/sitodt/modules/documenti_it/analisi_progammazione/finanziam_comunitari_ricerca/Final-Report.pdf; https://www.fondazionebrodolini.it/sites/default/files/technical_note.pdf.

[10] For a detailed description of the programme and the full list of available archives, see https://www.inps.it/it/it/dati-e-bilanci/attivit--di-ricerca/programma-visitinps-scholars.html.

(iii) unemployment-benefit and income-support archives; (iv) registers for collaborators, self-employed, and public-sector employees; and (v) workers' occupational and educational data derived from the *Mandatory Communications* database.[11]

Selected researchers work within the secure environment for limited periods (typically up to two years) under strict data-protection and disclosure rules. Researchers cannot extract individual-level data but may export aggregated, anonymised results subject to approval. They may also request the merging of project-specific external data with existing INPS archives, conditional on approval and a non-identifiability check conducted by INPS staff.

The *VisitINPS Scholars Programme* represents one of Italy's most ambitious initiatives to grant continuous research access to comprehensive labour market data. In its first three editions, 76, 99, and 92 projects were submitted—about half approved.[12] While acceptance rates are unavailable for later years, project numbers granted declined to roughly 32 per edition after 2020.[13] A non-exhaustive list of studies conducted within the programme includes Bamieh et al. (2025) on business uncertainty and labour flows; Casarico and Lattanzio (2024) on gender wage gaps; Ardito et al. (2024) on inequality in longevity; Fanfani (2023) on collective wage bargaining; and Mazzaferro (2023) on pension reform effects.

3 Collaborative Initiatives

The introduction of the *Mandatory Communications* data source and the political support for studies on the social determinants of health and inequality have stimulated several initiatives that promote the use of administrative data on employment, welfare, education and health for research purposes. In the sections that follow, we provide a review of some of these experiences and add some examples of both the use and integration of different administrative data sources.

3.1 Studies Based on the Mandatory Communications *Archive*

The Italian Ministry of Welfare established the Mandatory Communications online data collection system ("Comunicazioni Obbligatorie," hereafter *COB*) with the 2007 Budget Law. Since then, all employers in Italy have been required to use it to report every activation, transformation, and termination of employment contracts between

[11] For more information on the *Mandatory Communications* database, see § 3.1 in this chapter.

[12] https://www.inps.it/content/dam/inps-site/pdf/dati-analisi-bilanci/attivita-ricerca/visitinps-scholars/3376KEY-programma_ricerca_visitinps_scholars.pps.

[13] https://www.inps.it/it/it/dati-e-bilanci/attivit--di-ricerca/programma-visitinps-scholars/progetti-assegnati.html.

a worker and their company. The peculiarity of these data is their flow character because only "events" are recorded, while the stock of workers already active in the labour market in the years covered by the data, who did not experience any "event" at all (contract termination or transformation), remain invisible.

The COB data contains very rich information, with attributes on the worker (basic demographic data plus the level of education), very detailed information about the job, including the type of employment contract, the collective bargaining agreement used, a very detailed occupational classification (the Italian version of the ESCO classification at 6 digits) and the sector of activity (the Italian version of the NACE classification at 5 digits).

The availability of timely, population-based information has enabled a surge in studies on the Italian labour market, especially at the regional level. Many Italian regions now permit scientific access to their regional COB archives, which has made it possible to produce detailed, up-to-date analyses of local labour-market dynamics. At the national level, apart from the CiCo sample, mentioned previously (see above, paragraph 2.1), some collaborative initiatives involved the National Institute for Public Policies Analysis (INAPP), which granted access to a richer version of the database within its premises, based on *ad-hoc* agreements with requesting researchers.

At the local level, the Veneto Region was one of the first to exploit this new source, becoming a reference in the scientific community for the use of this kind of data for research purposes, thereby making an important contribution to best practices about decoding classifications and constructing relevant labour market indicators (Anastasia et al., 2009, 2016). They also developed a standardised database for research, a Public Use File (PUF) called "Mercurio," that contained detailed information built based on COB data.[14]

In other regions, the dissemination of COB data is not as standardised as in the case of the Veneto Region, but it is a common practice the signing of *ad-hoc* research agreements between universities and the regional institutions holding the data. In terms of drawbacks, this represents a stricter policy with respect to the case of Veneto's PUF or the national MFR version of the archive. In terms of advantages, the rationale of microdata dissemination by regional institutions is to advance research on subjects which are relevant for their institutional mandates and priorities. Consequently, the *ad-hoc* agreements have often a broader scope, including additional set of administrative data sources with respect to the COB (*e.g.*, data on active labour market provisions delivered by the same regional institution), to pursue a common research interest between the data owners and the researchers.[15] It is also noteworthy that, under the current personal data regulations, the fact that the data provider has

[14] https://www.venetolavoro.it/public-use-file.

[15] It is beyond the scope of this work to review all regional policies in terms of data access to support our statement. In Italy, the management of public employment services and the implementation of active labour market policies are the direct responsibility of the regions, which have a large degree of autonomy in shaping the regulatory system. As an illustration, we might quote the Piedmont Region, where the statute of the regional agency that manages the employment services (the Agenzia Piemonte Lavoro, APL) clearly states the relevance of research activities for the improvement of

an institutional interest in organising the data needed for the research also facilitates the record linkage to other individual databases.

The rise in collaborative initiatives is evident from the many scientific contributions published in the last decades across various Italian regions. A non-exhaustive list of examples includes: for Lombardy, Cornaggia et al. (2017) and Moscone et al. (2016), who both used COB data linked to drug prescriptions to study the impact of precarious work experiences on mental health; Sestito and Viviano (2018) and Ardito et al. (2023), who employed COB data from Veneto and Piedmont, respectively, to analyse the effects of hiring subsidies and dismissal costs regulations on the labour market; Cattani et al. (2023), who studied the association between contract type and job content using COB data from the Emilia-Romagna; and Pastore and Pompili (2020), who drew on COB data from Friuli Venezia Giulia to assess the impact of a regionally integrated programme of active labour policies.

3.2 National Studies on Health

The 2007–2009 economic crisis accentuated social inequalities across several dimensions, including health. As a result, research agendas aimed at studying social inequalities in a broader sense have investigated the role played by non-sanitary policies in shaping population health and have grown rapidly, thereby gaining a high political attention worldwide (Costa et al., 2014; Mackenbach et al., 2008; Marmot, 2005).

To study the social determinants of health, data containing both health and socioeconomic information are essential. Yet, in Italy, few existing databases were suitable for this purpose. To fill this gap, two national initiatives were launched, creating the Italian Longitudinal Study (ILS), linking survey data with national administrative health registers, and the Work and Health Italian Panel (WHIP), linking social security data with administrative health national registers.[16]

The ILS was created by linking National Health Survey (NHS) data from nationally representative samples of Italian citizens to long-term administrative health records. The project is managed by ISTAT, the Ministry of Health and the Region of Piedmont, with operational responsibility entrusted to the Piedmont Epidemiology Service (EpiPiemonte). Currently, the ILS is listed in the National Statistical Program with code IST-02566 (ISTAT 2019).[17]

---

their employment services and formal agreements with other institutions to achieve this objective (Resolution of the Regional Council 25/3/2019, n. 366-6856, article 1).

[16] The acronym of the original sample stood for Work Histories Italian Panel (see Footnote 1 above). Currently, the database is either dubbed WHIP-Salute (WHIP&Health), to stress its extension with a health data section, or simply WHIP, meaning Work and Health Italian Panel.

[17] The National Statistical Programme is the main normative act in Italy about statistical data production and access, was set up within the SISTAN (the National Statistical System) and is supervised by the Italian Guarantor for the Protection of Personal Data. For a discussion of privacy issues in epidemiological studies and public health surveillance, see Bisceglia et al. (2023).

Two databases have so far been released under this initiative, integrating information from the NHS conducted in 2000 and 2005 with administrative data on hospitalisations and deaths up to 2014. The NHS already includes cross-sectional information on self-reported health and basic socio-economic characteristics; its integration with administrative data adds certified medical information and a longitudinal dimension, enabling long-term follow-up to measure mortality and the hospitalisation risks. While the NHS is designed to be nationally representative, the sample size was enlarged from 24,000 households in the 2000 wave, to about 60,000 in 2005, in order to provide enough size for sub-regional health planning.

Data on mortality come from ISTAT's "Deaths and Causes of Deaths" (DCD) register, the most comprehensive source of information about causes of death in Italy. It is population-based and records all deaths occurring in the Italian territory during a calendar year. The causes of death are based on the medical information contained in the death certificates and these are coded according to the ICD-10, the International Statistical Classification of Diseases. The certificates identify the "initial cause of death" according to the WHO definition (i.e., "the illness or traumatism that initiated the chain of morbid events that led directly to death, or the set of circumstances of the accident or violence that caused the fatal traumatic injury").

Data on hospitalisations are derived from the Ministry of Health's Hospital Discharge Forms ("Schede di Dimissione Ospedaliera," SDO), the register of all hospitalisations in both public and private hospital facilities in Italy, including detailed information, among others, on the principal and secondary diagnoses (coded according to the ICD-10), the principal and secondary medical interventions delivered, and the Diagnosis-Related Group used to classify patients for the purposes of cost accountability.

In 2007, following a similar agenda, the National Centre for Disease Prevention and Control (CCM) funded a project promoted by EpiPiemonte and the University of Turin, the aim of which was to integrate the already existing WHIP sample derived from INPS data with information on occupational accidents and diseases held by the Italian Workers Compensation Authority (INAIL). Other institutions have joined the initiative over the years and have further extended the information about work and health biographies recorded in the sample: the ISTAT (causes of death), the Ministry of Health (Hospitalisations) and the Ministry of Welfare (Mandatory Communications). The database is managed by the Ministry of Health, is listed in the National Statistical Programme (code SAL-00054) and has been identified as a surveillance system of national relevance (PM Decree 3/3/2017).

The WHIP database has a particularly rich section on work biographies, integrating the INPS and COB data mentioned above for a national representative systematic sample of Italian working and retired population with a sampling rate of about 7%. The population coverage is almost complete: the only workers excluded are those belonging to categories of high professionals (such as architects and notaries) whose social security contributions are not managed by the INPS. However, the most detailed information is only available for employees in the private sector, while the information collected on public employees, quasi-dependent workers

and self-employed workers is limited to the length of employment spells and the corresponding earnings.

The section on health biographies includes, in addition to what is already available in the INPS data (such as sick leave periods and the receipt of disability benefits), data from the causes of death and hospitalisations described above, plus INAIL data on occupational diseases and accidents. The latter are very detailed, including (among others) the date and time of the accident, the part of the body affected, the type and severity of the injury and the time required for recovery. The INPS-INAIL linkage has allowed for the first systematic studies of occupational risk to be undertaken in Italy, given that the INAIL administrative source alone only records data on events (the injuries) and not on those exposed to the risk (the workers, see, e.g., Bena et al., 2012, 2013). The linkage between work biographies and occupational injuries, moreover, made it possible to study how critical periods in work careers, such as firm closures or labour market entry, and institutional factors such as the level of employment protection legislation are related to the risk of injury and to the return-to-work patterns (see, *e.g.*, Galizzi et al., 2019, 2023; Leombruni et al., 2013, 2019).

WHIP dynamic population comprises about 10 million people, tracked from the mid-seventies of the last century for matters concerning employment and retirement biographies. The balanced panel with both work and health biographies is shorter, given that INAIL data starts in 1994, while hospitalisations are available from 2001. The detailed data on occupations and educational attainment from the COB source are, unfortunately, available only for a small subset of the sample, due to the flow nature of the data source and its relatively recent start date.

Currently, the use of the ILS and the WHIP database for research purposes is restricted to the members of promoting institutions; therefore, data access for other researchers is limited to collaborative projects with the University of Turin and/or EpiPiemonte and to their premises. Regarding WHIP, however, the situation will change in the very near future, since the Ministry of Health included the possibility of disseminating an MFR version of the database in the last revision of the National Statistical Programme; the MFR has already been prototyped and agreed upon with the participating institutions.

3.3 The Italian System of Longitudinal Metropolitan Studies

Record linkage systems to investigate health and its social determinants and consequences have also been set up at the local level. The first pioneering study of this kind was conducted in Turin using data from the 1971 census and was carried out by EpiPiemonte in collaboration with the Municipality of Torino (Stringhini et al., 2015). At present, the Turin Longitudinal Study (SLT) consists of all available decennial censuses from the city (from 1971 to 2021), to which information from a wide range of health registers is linked at the individual level: this includes the resident register

of the Municipality of Torino, ISTAT's causes of death, hospitalisations, pharmaceutical prescriptions and specialised and emergency services, as well as other regional health databases such as the Piedmont Cancer Registry or the register of diabetic patients (Costa e Demaria, 1988; Costa et al., 2017).

Censuses provide a wealth of information on demographic and socio-economic indicators, both in terms of individual and household data (composite indicators, such as household type, housing type or social class) and in terms of statistics on the geographical area of residence (summary indicators of the socio-economic characteristics of the census section to which they belong). The study now comprises about 2.4 million individuals and is also included in the National Statistical Program (NSP) (PIE-00001).

Following in Torino's example, other census-based longitudinal metropolitan studies have been implemented that link censuses to mortality and morbidity registers (Caranci et al., 2018). These include the Tuscan Longitudinal Study, covering the cities of Livorno from 1981, Firenze since 1991 and Prato since 2001 (Biggeri et al., 1999); the Emilian Longitudinal Study, covering Reggio Emilia since 1991 and Bologna and Modena since 2001 (Candela et al., 2005); and the Longitudinal Studies of Venezia and Roma, both launched with the 2001 census (Cacciani et al., 2015; Simonato et al., 2009).

These local initiatives have converged into a formally established national system, the Italian Network of Longitudinal Metropolitan Studies (IN-LiMeS, Caranci et al., 2018). Its goal is to standardise regional cohort studies in order to create a robust, coherent and nationally extensive repository of integrated data on health outcomes and mortality, together with demographic and socioeconomic information. The minimum core dataset required to participate in the IN-LiMeS network includes municipal resident registers, census data, mortality registers and hospitalisation archives, all linked at the individual level through deterministic record linkage. Confidentiality is ensured by pseudo-anonymisation (i.e., removing only direct personal identifiers), which preserves the full research potential of the data. Access to external researchers is not granted a priori, but data owners have a long-standing tradition of collaboration with universities, formalised by *ad-hoc* agreements to carry out projects with internal staff.

In addition to these metropolitan longitudinal studies, numerous regional initiatives exist in which researchers collaborate directly with regional health administrative offices through ad-hoc agreements for specific projects. These collaborations provide access to detailed administrative health registers and have enabled rigorous analyses of diverse topics, including the social determinants of health, healthcare utilisation patterns, and the costs associated with different services (Atella et al., 2006; Bruni et al., 2016; Di Giacomo et al., 2025). While these projects are often limited in scope and duration, they underscore the substantial, yet largely untapped, potential of regional administrative health data for producing high-quality research that addresses pressing policy questions.

4 Initiatives on Education Data

Education is a crucial dimension for the study of social, health and economic outcomes, yet information on education in Italian data sources is often limited. For example, in most of the sources mentioned so far the information on education is sometimes absent (as in the data coming from the Social Security Administration), sometimes recorded with a poor level of detail (as in the case of self-reported survey data) and sometimes with severe data quality issues (as in the COB source, in which there is a high percentage of missing and inconsistent data). Linking survey and administrative data with education records managed by public institutions involved in the education system can provide much more accurate and granular information, enabling researchers to study education trajectories and their links to labour market and social outcomes.

A first important source is the data collected and distributed by the National Institute for the Evaluation of the Educational and Training System (INVALSI), which manages a mandatory, population-based national survey on students' learning for all educational levels up to the upper secondary level (the "Test INVALSI"). These dataare, in fact, easily accessible and are freely disseminated as Open-Data. Furthermore, fully population-wide anonymised microdata are also available to researchers belonging to recognised research institutions and shall be issued upon registration and filling in a specific form.[18] INVALSI data have already been extensively used in empirical research: for example, to study how classroom composition affects student outcomes in southern Italy (Angrist et al., 2017) and to investigate the role of stereotypes in shaping academic performance among immigrant students (Alesina et al., 2024).

Although collected and stored at the relevant ministries' statistical offices, administrative data on students' careers (enrolments, grades, school leaving certificates) up to the upper secondary level of education are less accessible for researchers, compared to university-level education data. This is simply because studies using data from the researchers' own universities can be carried out with fewer barriers regarding personal data exchange. However, this "inside" access does not generally cover the mobility of students between universities or their subsequent employment careers. Actually, national-level university careers are collected in a very harmonised way, thanks to the National Register of Students (Anagrafe Nazionale degli Studenti e dei Laureati, ANS), that collects nation-wide, harmonised data on university careers in Italy,[19] but there is currently no standardised procedure to grant access to it, in the spirit of the ones quoted above about the CiCo or the MFR samples.

Several initiatives have emerged to address these limitations. The AlmaLaurea Interuniversity Consortium[20] manages surveys administered to all students and graduates across all stages of their education, allowing the match of administrative university records with, for example, five-year follow-ups of students' transitions into the

[18] https://serviziostatistico.invalsi.it/en/catalogo-dati/.

[19] See https://www.miur.gov.it/anagrafe-nazionale-degli-studenti-e-dei-laureati.

[20] https://www.almalaurea.it/en/.

labour market. A drawback of this system is that only the students at a given university can have their survey matched to administrative records from that university. The national database, on the contrary, is not accessible to researchers outside the AlmaLaurea Consortium through any standardised procedure.

The "co-location" of the research community and the administrative data greatly enhances the possibility of initiatives in which different universities set up a collaborative network to share their data under a common research interest. An interesting initiative in this respect is the "Mobysu.it" database, concerning the mobility of Italian university students. It was developed through a network of universities and national public administrations, enabling access to university career data at the national level, matched with AlmaLaurea and INVALSI information (see, *e.g.,* Priulla et al., 2023; Rizzi et al., 2021; Usala et al., 2024).

Although AlmaLaurea provides valuable information on graduates' employment outcomes, the potential would be greatly enhanced by systematic linkage with administrative labour market data. To date, no national-level systematic activity links university data with Mandatory Communications or INPS data. Nevertheless, some universities have pioneered such projects: e.g., project Uni.CO at Sapienza University of Rome (see Alleva et al, 2015; also applied to Luiss Guido Carli University data); SKiNT at the University of Turin (project SKiNT; see Di Monaco et al, 2019); and the Milan Higher Education Observatory (MHEO) at the University of Milano (see Mezzanzanica & Salini, 2025). Another promising source of matched university-level data is the *EDUSCOPIO* project, which uses administrative records from the *National Archive of Students and Graduates* (ANS), linked to students' labour market outcomes after graduating from high school (Aina et al., 2021).

5 Conclusions

The story that we have told about the use of administrative data for scientific research in Italy is one of both excellences and contradictions. One of the main turning points, which has largely shaped the current situation, has involved the introduction of a centralised policy by the Ministry of Welfare for the dissemination of national samples of INPS and COB data and the INPS visiting scholars' programme. One advantage of this setting concerns data accessibility, which is potentially granted to any researcher for free, at least in the case of the Micro-Files for Research, while in the case of the INPS visiting programme, because of rationing, the current system is potentially prone to gatekeeping.

Conversely, dysfunctional features remain in place. Researchers have access to the administrative sources without the statistical treatment and metadata preparation that is necessary for their use for research purposes (Wallgren & Wallgren, 2007). Regarding metadata, while traditional surveys are based on questionnaires distributed together with the data, administrative information is collected as a result of a (usually very complicated) set of norms and regulations that are difficult to reconstruct for an individual researcher. In this respect, current access to data in the centralised setting

is characterised by very poor documentation and no systematic production of the metadata necessary for research purposes. The lack of a statistical pre-processing of the data, moreover, poses serious risks of misidentification of statistical units, particularly for businesses, and the use of poor attribute definitions, especially for the identification of employment relationships and their duration (see Benedetto et al, 2007; Contini et al, 2009; Hethey-Maier & Schmieder, 2013; Leombruni, 2024; US BoLS, 2022, for a wider discussion on these points). The current management of data in the centralised setting then is far from the best practices followed where the centralised policy is managed by national statistical institutes, such as in the Nordic European Countries (see again Wallgren & Wallgren, 2007, and the chapters about Denmark, Norway and Sweden in this book).

The collaborative setting almost mirrors the situation of the centralised one. On the *downside*, the access to the data is characterised by an insider/outsider mechanism, with consequences for transparency, ease of access and the replicability of processes. Conversely, direct collaboration between data owners and the research community typically allows for a better definition of data requirements and the implementation of all the data transformation and documentation necessary to make the administrative data suitable for scientific purposes.

We envisage two main directions for a better use of administrative data for research in Italy. The first concerns a greater effort on data integration. We have already mentioned the research potential that would be achieved by integrating administrative data on education into existing databases. We also observed how the COB data's availability has contributed to a surge of studies, but following a data-driven route means that much of the evidence on the labour market, on which regional planning is currently based, considers flow-data only. This implies that workers already in employment prior to the introduction of the Mandatory Communications system are an almost invisible plateau to the regional policy maker. The integration of COB flow data with the stock data available in the INPS archives would be a straightforward solution to this problem. Indeed, data integration may be challenging, but it is not true, as is sometimes held, that the main obstacle here regards the data protection regulations. The several examples that we have provided, about linking health registers to occupational data, are testament to the fact that current regulations—while strict—do allow for these kinds of projects.

A second direction of improvement concerns data accessibility. A major improvement would be achieved if more significant efforts were made in increasing and standardising data accessibility practices. In both the centralised and in the collaborative settings, enhancements could be achieved by trying to build on each other's strengths (*e.g.*, by trying to capitalise on the experiences accrued in the scientific community to pre-process and document the data distributed in the centralised setting) or by making cooperation with outsiders more transparent, with the introduction of systematic visiting programmes and through the systematic production of Micro File for Research and Public Use File versions of the databases. Further improvements could also be achieved with technological solutions, such as a larger use of remote access and remote execution: while the first is not always allowed by current regulations, the latter can always be implemented by distributing mock files in order to test the

data analysis pipelines. A further, promising perspectives that is worth mentioning in this respect is the so-called federated learning, which allows for a decentralised analysis of data so that a research project can be implemented as though it had access to full data without the actual distribution of micro data sources.

References

Acciari, P., Alvaredo, F., & Morelli, F. (2024). The concentration of personal wealth in Italy 1995–2016. *Journal of the European Economic Association, 22*(3), 1228–1274.

Aina, C., Bratti, M., & Lippo, E. (2021). Ranking high schools using university student performance in Italy. *Economia Politica, 38*(1), 293–321.

Alesina, A., Carlana, M., La Ferrara, E., & Pinotti, P. (2024). Revealing stereotypes: Evidence from immigrants in schools. *American Economic Review, 114*(7), 1916–1948.

Alleva, G., Magni, C., Lucisano, P., Renda, E., & Petrarca, F. (2015). La domanda di lavoro per i laureati. I risultati dell'integrazione tra gli archivi amministrativi dell'Università sapienza di Roma e del Ministero del lavoro e delle politiche sociali. Roma: *Ed. Nuova Cultura.*

Anastasia, B., Disarò, M., Gambuzza, M., & Rasera, M. (2009), Comunicazioni obbligatorie e analisi congiunturale del mercato del lavoro: evoluzione, problemi metodologici, risultati, *Veneto Lavoro,* collana "I Tartufi", n. 35

Anastasia, B., Emireni, G., Gambuzza, M., Maschio, S., & Rasera, M. (2016). Grammatica Delle Comunicazioni Obbligatorie/3. Guida alle elaborazioni a partire dai dati di flusso. *Veneto Lavoro,* collana Tempi & Metodi, marzo 2016

Angrist, J. D., Battistin, E., & Vuri, D. (2017). In a small moment: Class size and moral hazard in the Italian Mezzogiorno. *American Economic Journal: Applied Economics, 9*(4), 216–249.

Ardito, C., Berton, F., & Pacelli, L. (2023). Combined and distributional effects of EPL reduction and hiring incentives: An assessment using the Italian "Jobs Act." *The Journal of Economic Inequality, 21*(4), 925–954.

Ardito, C., Zengarini, N., Leombruni, R., Costa, G., & d'Errico, A. (2024). Increasing inequalities in longevity among Italian workers. *Oxford Economic Papers, 76*(4), 1128–1146.

Atella, V., Peracchi, F., Depalo, D., & Rossetti, C. (2006). Drug compliance, co-payment and health outcomes: Evidence from a panel of Italian patients. *Health Economics, 15*(9), 875–892.

Bamieh, O., Coviello, D., Ichino, A., & Persico, N. (2025). Effect of business uncertainty on turnover. *Journal of Labor Economics, 43*(2), 319–348.

Bena, A., Giraudo, M., Leombruni, R., & Costa, G. (2013). Job tenure and work injuries: A multivariate analysis of the relation with previous experience and differences by age. *BMC Public Health, 13*, 1–9.

Bena, A., Leombruni, R., Giraudo, M., & Costa, G. (2012). A new Italian surveillance system for occupational injuries: Characteristics and initial results. *American Journal of Industrial Medicine, 55*(7), 584–592.

Benedetto, G., Haltiwanger, J., Lane, J., & McKinney, K. (2007). Using worker flows to measure firm dynamics. *Journal of Business & Economic Statistics, 25*(3), 299–313.

Berta, P., Lovaglio, P. G., & Verzillo, S. (2021). How have casemix, cost and hospital stay of inpatients in the last year of life changed over the past decade? Evidence from Italy. *Health Policy, 125*(8), 1031–1039.

Berton, F., Pacelli, L., Quaranta, R., & Trentini, F. (2023). Patterns of labor market reforms: A regional approach to the Italian 'Jobs Act.' *Sinappsi, 13*, 50–67.

Biggeri, A., Marchi, M., Dreassi, E., Baldi, P., Benvenuti, A., & Merler, E. (1999, July–September). Studio Longitudinale Toscano: un'analisi per quartiere per le città di Firenze e Livorno [The Tuscany longitudinal study: mortality among selected causes in inner city of Florence and Leghorn]. *Epidemiol Prev., 23*(3), 161–74.

Bisceglia, L., Caranci, N., & Zengarini, N. (2023). Privacy and epidemiology: Let's find a shared solution. *Recenti Progressi in Medicina, 114*(6), 332–336.
Brescianini, S., Cappellari, L., & Checchi, D. (2024). *Heritability in the labour market: Evidence from Italian twins* (IZA Discussion Paper No. 17476). IZA – Institute of Labor Economics.
Bruni, M. L., Mammi, I., & Ugolini, C. (2016). Does the extension of primary care practice opening hours reduce the use of emergency services? *Journal of Health Economics, 50*, 144–155.
Cacciani, L., et al. (2015). Education and mortality in the Rome longitudinal study. *PLoS One*, *10*.
Candela, S., Cavuto, S., & Luberto, F. (2005). *Condizioni socioeconomiche e mortalità nelle popolazione di Reggio Emilia.* Azienda USL di Reggio Emilia.
Caranci, N., Di Girolamo, C., et al. (2018). Cohort profile: The Italian Network of Longitudinal Metropolitan Studies (IN-LiMeS), a multicentre cohort for socioeconomic inequalities in health monitoring. *BMJ Open, 8.*
Casarico, A., & Lattanzio, S. (2023). Behind the child penalty: Understanding what contributes to the labour market costs of motherhood. *Journal of Population Economics, 36*(3), 1489–1511.
Casarico, A., & Lattanzio, S. (2024). What firms do: Gender inequality in linked employer-employee data. *Journal of Labor Economics, 42*(2), 325–355.
Cattani, L., Dughera, S., & Landini, F. (2023). Interlocking complementarities between job design and labour contracts. *Italian Economic Journal, 9*(2), 501–528.
Contini, B., Leombruni, R., Pacelli, L., & Villosio, C. (2009). Wage mobility and dynamics in Italy in the 1990s. In E. P. Lazear & K. L. Shaw (Eds.), *The structure of wages: An international comparison* (pp. 373–400). University of Chicago Press.
Contini, B., & Revelli, R. (1987). The Process of Job Creation and Job Destruction in the Italian Economy. *Labour, 1*(3), 121–144. https://doi.org/10.1111/j.1467-9914.1987.tb00122.x
Contini, B., & Revelli, R. (1989). The relationship between firm growth and labor demand. *Small Business Economics, 1*, 309–314.
Conti, R., Bavaro, M., Boscolo, S., Fabrizi, E., Puccioni, C., Ricchi, O., & Tedeschi, S. (2023). The Italian Treasury Dynamic Microsimulation Model (T-DYMM): Data, structure and baseline results. *Government of the Italian Republic (Italy), Ministry of Economy and Finance, Department of the Treasury Working Paper* (1).
Cornaggia, C. M., Beghi, M., Mezzanzanica, M., Ronzoni, G., Vittadini, G., & Maffenini, W. (2017). Psychotropic drug consumption and employment status in time of economic crisis (2007–2011). *Psychiatric Quarterly, 88*(2), 371–384.
Costa, G., Bassi, M., Gensini, G. F., Marra, M., Nicelli, A. L., & Zengarini, N. (Eds.). (2014). *L'Equità nella salute in Italia.* Franco Angeli Editore.
Costa, G., & Demaria, M. (1988). Un sistema longitudinale di sorveglianza della mortalità secondo le caratteristiche socio-economiche, come rilevate ai censimenti di popolazione: descrizione e documentazione del sistema. *Epidemiologia e prevenzione, 36*, 37–47.
Costa, G., Stroscia, M., Zengarini, N., & Demaria, M. (2017). *40 anni di salute a Torino.* Spunti per leggere i bisogni e i risultati delle politiche, Inferenze, Milano.
Di Giacomo, M., Piacenza, M., Salmasi, L., & Turati, G. (2025). Understanding labour productivity in maternity wards. *Labour Economics*, 102760.
Di Monaco, R., Leombruni, R., Mosca, M., & Pilutti, S. (2019). Il disallineamento tra laurea e occupazione. Uno studio per nuovi indici di qualità del match dei laureati nel mercato del lavoro italiano. In S. Scamuzzi (Ed.), *Studenti universitari: ingresso, carriera, esito professionale.* Università di Torino.
Di Zio, M., Falorsi, S., & Loriga, L. (2022). Overview of the Istat activities and open problems. In *First workshop on methodologies for official statistics—Proceedings. Session. Vol. 1.* ISTAT.
Fanfani, B. (2023). The employment effects of collective wage bargaining. *Journal of Public Economics, 227*, Article 105006.
Galizzi, M., Leombruni, R., & Pacelli, L. (2019). Successful return to work during labor market liberalization: The case of Italian injured workers. *Journal of Labour Market Research, 53*, 9.
Galizzi, M., Leombruni, R., & Pacelli, L. (2023). Severe work disabilities and long-lasting losses. *Labour Economics, 85*(C).

Leombruni, R. (2024). Interviewing Administrative Records: A Conceptual Map for the Use of Big Data for Economic Research. *Statistica Applicata - Italian Journal of Applied Statistics, 36*(3), 295–326.

Leombruni, R., Razzolini, T., & Serti, F. (2013). The pecuniary and non-pecuniary costs of job displacement. The risky job of being back to work. *European Economic Review, 61*, 205–216.

Leombruni, R., Razzolini, T. & Serti, F. (2019). Macroeconomic conditions at entry and injury risk in the workplace. *Scandinavian Journal of Economics, 121*.

Hethey-Maier, T., & Schmieder, J. F. (2013). Does the use of worker flows improve the analysis of establishment turnover? Evidence from German administrative data. *Journal of Contextual Economics-Schmollers Jahrbuch, 4*, 477–510.

Mackenbach, J. P., Stirbu, I., Roskam, A. J., Schaap, M., Menvielle, G., Leinsalu, M., & Kunst, A. (2008). Socioeconomic inequalities in health in 22 European countries. *New England Journal of Medicine, 358*, 2468–2481.

Marmot, M. (2005). Social determinants of health inequalities. *The Lancet, 365*(9464), 1099–1104.

Mazzaferro, C. (2023). The transition to NDC in Italy: Assessing distributive and financial effects. *Journal of Pension Economics & Finance, 22*(1), 116–131.

Mezzanzanica, M., & Salini, S. (2025). *III Rapporto MHEO - Dopo la laurea: gli sbocchi occupazionali dei laureati milanesi e lombardi*. MHEO - Milan Higher Education Observatory.

Moscone, F., Tosetti, E., & Vittadini, G. (2016). The impact of precarious employment on mental health: The case of Italy. *Social Science & Medicine, 158*, 86–95.

Pastore, F., & Pompili, M. (2020). Assessing the impact of off-the-job and on-the-job training on employment outcomes: A counterfactual evaluation of the PIPOL program. *Evaluation Review, 44*(2–3), 145–184.

Pinotti, P. (2017). Clicking on heaven's door: The effect of immigrant legalization on crime. *American Economic Review, 107*(1), 138–168.

Priulla, A., Vittorietti, M., & Attanasio, M. (2023, October 1). Does Taking Additional Maths Classes in High School Affect Academic Outcomes? *Socio-Economic Planning Sciences, 89*, 101674.

Raitano, M. (2012). La distribuzione delle retribuzioni da lavoro dipendente in Italia nel periodo 1996–2009: l'evidenza dal campione AD-SILC. *Economia & lavoro, 46*(3), 9–28.

Rizzi, L., Grassetti, L., & Attanasio, M. (2021, June 21). Moving from North to North: How are the students' university flows? *Genus, 77*(1), 8.

Sestito, P., & Viviano, E. (2018). Firing costs and firm hiring: Evidence from an Italian reform. *Economic Policy, 33*(93), 101–130.

Simonato, L., Canova, C., & Tessari, R. (2009). *La geografia delle malattie nel comune di Venezia* (pp. 1–204). Cleup.

Stringhini, S., Spadea, T., Stroscia, M., Onorati, R., Demaria, M., Zengarini, N., & Costa, G. (2015). Decreasing educational differences in mortality over 40 years: Evidence from the Turin Longitudinal Study (Italy). *Journal of Epidemiology and Community Health, 69*(12), 1208–1216. https://doi.org/10.1136/jech-2015-205673

Usala, C., Sulis, I., Salaris, L., Tedesco, N., Porcu, M., & Barbieri, B. (2024). Disentangling the secondary school effects in the assessment of university inefficiencies. *Studies in Higher Education*, 1–21.

US Bols, US Bureau of Labor Statistics. (2022). s.d. *Employee Tenure Technical Note - 2022 A01 Results*. https://www.bls.gov/news.release/tenure.tn.htm. Accessed 10 August 2023.

Wallgren, A., & Wallgren, B. (2007). *Register-based statistics: Administrative data for statistical purposes*. John Wiley & Sons.

Roberto Leombruni is Associate Professor of Economic Statistics at the University of Torino, where he teaches Econometrics, Simulation models and Public Policy Evaluation. He is also Senior Researcher at the LABORatorio Revelli—Centre for Employment Studies and founder of the academic spin-off Metro-Polis srl. He coordinates, on behalf of the Italian Ministry of

Health, the development of WHIP-Salute, one of the largest databases for the study of work and health biographies in Italy. His research interests concern the multidirectional relations between health, work careers, retirement and welfare policies, the use of Big Data and ML/AI for scientific research, the use of Agent-Based models and Microsimulation Models for ex-ante policy evaluation, with contributions published in several international journals among which the European Economic Review, Labour Economics, BMC Public Health, The Scandinavian Journal of Economics.

Chiara Ardito , Ph.D. in Economics, is Quantitative Policy Analyst at the Joint Research Centre of the European Commission. She is an applied microeconomist whose research interests include policy impact evaluation using non-experimental methods, health economics, labour economics, the relationship between work, health and social inequalities in the context of ageing populations.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 17
Administrative Data Access in Latvia



Ieva Bloma, Jānis Glazkovs, and Fabio Berton

Abstract This chapter explores the role of administrative data in Latvia, where the Law on Statistics defines administrative registers as key sources for state data production. The chapter focuses on three applications: census, healthcare, and deportations. In the past three decades, census methodologies have evolved from an approach centred on interviews (and complemented with administrative registers) to an automated machine-learning approach based entirely on administrative registers and where interviews have been abandoned. In healthcare, Latvia has advanced digitalisation through e-Health, interoperability initiatives, and the establishment of the Latvian Digital Health Centre. Legislative innovations, including the Biobank Law, further support secure data use and patient-centred digital governance. Finally, the chapter suggests that the usage of administrative data in constructing historical memory and shaping national identity may also have positive spillovers into other fields. Overall, the future looks bright following the intensive integration and use of administrative data. Existing projects still require time to fully deploy their effects and tight cooperation among a variety of actors, though.

I. Bloma
European Union Agency for Asylum, Valletta, Malta
e-mail: Ieva.Bloma@euaa.europa.eu

J. Glazkovs
Permanent Representation of Latvia to the European Union, Brussels, Belgium
e-mail: janis.glazkovs@mfa.gov.lv

F. Berton (✉)
European Commission Joint Research Centre, Ispra, Italy
e-mail: fabio.berton@ec.europa.eu

© The European Union and The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_17

1 Introduction

The main legal act defining the role of administrative data at the state level, within the official state statistics' ecosystem, is the Law on Statistics, which envisages that "*one of the data sources of the official statistics is administrative data sources*".[1] Among other things, the Law on Statistics regulates the dissemination of data, and the dissemination of such data that would allow direct or indirect identification of a private individual in particular. The usage of the latter is allowed for research purposes—i.e. such "*work that provides case studies, analysis, drawing up of conceptual proposals or impact assessment* [within the policy planning process]".[2]

The Law on Statistics also stipulates that the Central Statistical Bureau of the Republic of Latvia (hereinafter—the Central Statistical Bureau), which operates under the Ministry of Economics of the Republic of Latvia's direct supervision,[3] is the main governmental authority that coordinates the preparation, collection, processing, and dissemination of the official state statistics for Latvia.[4] The Central Statistical Bureau has defined General Quality Requirements for the state authorities that provide official statistics in order to ensure the quality of the official statistics.[5] Among other things, these quality requirements outline the collection and usage of administrative data.[6] In particular, the guidelines emphasise the rationale for the usage of administrative data in order to reduce the burden on survey respondents, which in turn is to be used as a data collection method when (quality) administrative data are not available.

Access to administrative data for research is legally possible, but is not fully centralised. The Central Statistical Bureau serves as the main entry point, offering a formal application process and secure remote access to anonymised microdata, with metadata being available in English. Requests must be filed directly with the respective authority, typically assessed on a case-by-case basis, for sector-specific datasets (health, tax, education). There is no single standardised national procedure, although

[1] Law on Statistics, Art. 13 (1). The Law on Statistics was adopted on 4 June 2015 and entered into force on 1 January 2016 (with further amendments adopted on 24 October 2019 and 14 December 2023). An unofficial English translation of an incomplete version of the Law on Statistics (i.e. without the most recent amendments of 2023) is available here: https://likumi.lv/ta/en/en/id/274749-statistics-law.

[2] Law on Statistics, Art. 25 (4), unofficial English translation: https://likumi.lv/ta/en/en/id/274749-statistics-law.

[3] The website of the Ministry of Economics of the Republic of Latvia (in English): https://www.em.gov.lv/en.

[4] The website of the Central Statistical Bureau of the Republic of Latvia (in English): https://www.csp.gov.lv/en.

[5] More information on the quality assurance framework can be found here: https://www.csp.gov.lv/en/quality-assurance-framework However, the text of the General Quality Guidelines is only available in Latvian, here: https://www.csp.gov.lv/lv/visparigas-kvalitates-prasibas.

[6] Centrālās statistikas pārvaldes Statistikas nodrošināšanas kvalitātes vadlīnijas, v.2.0, 2023: https://www.csp.gov.lv/lv/visparigas-kvalitates-prasibas.

the Central Statistical Bureau provides the most structured path. The quality of documentation varies across institutions, and English language metadata are not always guaranteed outside the Central Statistical Bureau. The National Health Service is usually the first contact point, in relation to health data, and it might grant access, but only upon sustained request, often requiring legal agreements or ethical approval.

This chapter explores three case studies that depict the value generated by the usage of administrative data in Latvia, to public authorities on the one hand (with the case of official state statistics in the following Sect. 2) and to the general public on the other (with the cases of health data in Sect. 3 and of deportation data in Sect. 4). Section 5 will draw some tentative conclusions.

2 The Usage of Administrative Data in Official State Statistics: Census Data

The first official census after Latvia regained its independence from the Soviet Union (1990–1991) took place in the year 2000 and—even though the primary data collection method included interviews with respondents at their place of residence—this was also the first census in which some administrative data were used (more precisely, data from the Register of Natural Persons and data from the State Revenue Service). The following census took place in 2011, and it involved an increased usage of administrative data. In addition to the administrative data sources already used in 2000, the data from the State Land Service's information system on real estate (so-called Cadastre data) were also used. Interviewing still remained the primary data collection method employed. The data collected through interviews with respondents were further combined with administrative data from the registers and information systems mentioned previously.

The census of the year 2021 marked a critical turning point, however, as this census was based solely on administrative data and interviews were abolished entirely as a method for data collection. Essentially, the estimates were obtained by combining both the so-called signs of life approach and binary regression modelling.[7] Hence, the data from various registers and information systems allowed for the estimation of, for instance, the number of households in the country as well as the number, age, sex, and other characteristics of people belonging and related to each household. The administrative data sources used in 2021 included both the registers and information systems used already in 2000 and 2011—namely, the Register of Natural Persons, the State Revenue Service's data, the State Land Service's data on real estate, and new data sources that included registers maintained by governmental authorities (e.g.

[7] The "signs of life" approach in census data collection indicates a method to determine whether a housing unit is occupied without using direct evidence (e.g., an interview), while relying instead on indirect hints (cars parked outside, electricity consumption, etc.). Within this method, a binary dependent variable—where zero stands for vacant unit and one for occupied—is modelled as a function of a set of explanatory variables, in this case being drawn from administrative registers.

the State Employment Agency, the State Social Insurance Agency, the Ministry of Education), state- or municipality-owned enterprises (e.g. the Maritime Administration of Latvia which maintains the Register of Seafarers) as well as non-governmental associations (e.g. data provided by the Council of the Creative Unions of Latvia), and some private enterprises (e.g. the enterprises providing heating or water services).

The usage of administrative data exclusively in conducting the census was not entirely unproblematic. Much depends on both administrative data quality and the model(s) applied in carrying out the estimates. While the former was consistently improved through inter-institutional cooperation between the Central Statistical Bureau and the various organisations maintaining each specific administrative source, the shortcomings of the latter were particularly highlighted by the COVID-19 pandemic and by Russia's full-scale invasion of Ukraine. The binary regression model used in the census of 2021 is certainly particularly sensitive to swift changes in administrative data. However, disruptive events can rapidly change the structure of administrative data. Indeed, the COVID-19 pandemic heavily affected data regarding employment and the number of doctors' visits, while Russia's invasion of Ukraine affected migration flows as well as the procedure for how Russian citizens can obtain a temporary residence permit in Latvia. As a result, the structure of administrative data that was feeding the binary regression model used to implement the "signs of life" approach changed, and it became more difficult for the model to produce reliable results.

To tackle the shortcomings described above, the Central Statistical Bureau has started to apply unsupervised machine learning in carrying out population estimates since 2025. The new model still combines administrative data and the so-called signs of life of individuals, but instead of applying a supervised binary regression model, the administrative data are now analysed by applying unsupervised machine learning. According to the Director General of the Central Statistical Bureau, such a shift will allow for population estimates to be carried out annually, instead of once per decade (as was the case until now), while still eliminating the administrative burden for respondents and keeping the general costs of population estimates low.

3 Health Data

Over the past decade Latvia has undergone rapid digital transformation in the public sector, including in the healthcare sector. This development is increasingly impacting both patients and medical staff in their daily lives, as well as the overall efficiency and quality of the national healthcare system itself. From the e-Health system (since 2016, see below) and e-prescriptions to genomic data management solutions, technologies are gradually being introduced that have the potential to significantly improve diagnostic and treatment quality, as well as public health indicators. Latvia ranks among those countries with a relatively high level of digital maturity in healthcare services—77 points (the EU average is 70).

The digital data ecosystem in healthcare is complex, involving many independent institutions. It consists of various types of patient health data generated by medical institutions, which is stored in their information systems and partly in the e-Health system. Additional data include pharmacy medication records and lab test results that are processed and stored in their respective pharmacy and laboratory systems, some of which are also reflected in the e-Health system. In addition, healthcare data include administrative, statistical, monitoring, and supervisory information created by healthcare providers, pharmacies, educational and certification institutions, and health sector authorities and are stored in national health information systems. These include the e-Health system, disease registries, statistical databases, healthcare provider and professional registers, drug and medical device registers, reimbursement systems, and platforms for international data exchange. The digitalisation and integration of healthcare systems is an ongoing process, though.

To that end, this section examines health data management in Latvia from multiple perspectives—analysing user (patients and professionals) experiences, legal and structural development trends, and outlining future directions and anticipated changes in digitalisation and data governance.

3.1 Health Data Management from the Perspective of Patients

Health data is most commonly associated with the e-Health system (www.eveseliba.gov.lv) for Latvian residents, where records of doctor visits, e-prescriptions, sick leave certificates, vaccination data, and hospital discharge summaries are available. Users access the portal using *eParaksts* (eSignature) or SMART-ID. The portal contained over 277,000 hospital discharge summaries and nearly 6 million laboratory test result reports in 2024, according to the Latvian Digital Health Centre.

However, patients' experience remains fragmented and access to their personal data is often complicated. This relates technically to user interface or authentication issues and in terms of content comprehension. Patients frequently complain that the data displayed in the e-Health system is incomplete or outdated, while the timely transfer of information between doctors often depends on individual initiative or on the use of paper documents. In many cases, patients are responsible for delivering their medical documents between doctors, often in physical format. The data exists, but it is not always structured, understandable, or accessible without additional interventions. Moreover, different medical institutions use different information systems, and synchronisation does not occur automatically in most cases. Another significant concern is trust in data security. While Latvia has not experienced any major health data breaches to date, there is public concern about how data is used and how effectively it is protected.

3.2 *Health Data Management from the Perspective of Medical Institutions*

Healthcare data in Latvia remains fragmented and is only partially digitalised. Medical institutions do not generally have access to investigations and procedures carried out by other facilities, as data sharing between institutions is limited. Many healthcare providers use information systems that are not interoperable with the national e-Health system or that are capable of sending, but not receiving, data. There is a lack of unified data standards and a centralised approach. Moreover, the core health data, such as basic patient information and test results, are only partially available in the e-Health system. Radiological images are commonly not uploaded at all.

Another major issue is the weak data exchange between healthcare and social care services, which hinders the continuity of patient care. Some medical institutions have developed electronic health cards that allow patients to access their own data, but these are used inconsistently, often incurring a fee, and may result in inequality. Overall, the current system does not support reliable, convenient, or efficient data exchange, neither between institutions nor across borders, thereby hindering the quality of care and the sustainable development of the health system. In addition, doctors and nurses frequently face an administrative burden related to entering data into multiple systems. While the digitalisation of data is theoretically intended to reduce administrative burdens, in practice the lack of integration between information systems often results in parallel data entry into both internal systems and the national e-Health system. This consumes time and reduces any motivation to fully utilise digital tools.

A further challenge lies in the uneven cooperation between primary care providers and hospitals regarding data exchange. The availability and flow of information often depend upon regional infrastructure, information system compatibility, and interoperability, as well as legal/contractual arrangements. While digital solutions aim to reduce redundant documentation and improve data flows, achieving these goals requires more than just technical solutions. At the same time, medical professionals acknowledge that if data were accessible in a structured and reliable manner, then it would significantly support diagnostics, treatment planning, and long-term health analysis. Greater potential to use health data in research and personalised medicine is viewed as a future opportunity that requires a robust data governance infrastructure.

Both the volume and importance of health data are increasing significantly as medical approaches evolve and technology develops rapidly. The need for change is widely recognised—Latvia's digital health ecosystem is fragmented, lacks common data standards, and suffers from decentralised management of digital solutions which leads to challenges in data quality, interoperability, and effective governance.

3.3 Current Developments: Structural Reforms and Legislative Initiatives

The Public Health Guidelines for 2021–2027 and the Digital Transformation Guidelines for 2021–2027, along with the principal strategic document the Digital Health Strategy until 2029, define and guide the development of Latvia's health information systems, while ensuring the quality and security of data exchange. These documents are aligned with European Union regulations and Latvian legislation, which is essential for safeguarding patient rights and data privacy, as well as for the effective governance and analysis of the healthcare system. Together, they lay the foundation for a modern, secure, and user-centred health information environment in Latvia that supports both the quality of care and innovation.

By the end of 2024, the Ministry of Health oversaw 19 national information systems and 7 registers. However, the management of these systems had been decentralised across various sectoral institutions. Following consultations with key stakeholders, a new institution—the Latvian Digital Health Centre (LDHC)—was established at the end of 2024. It is responsible for the governance of health data at the national level. Its mission is to review and to modernise existing processes and to introduce unified standards and digital solutions. This approach shall enable data-driven policymaking and steer the healthcare system towards more results-oriented governance while enhancing care efficiency, safety, and user experience.

The LDHC is tasked with maintaining these systems, providing methodological leadership, developing, and updating data classifiers and facilitating Latvia's integration into the European Health Data Space (EHDS).[8] Over the next five years, Latvia and all EU Member States are expected to gradually harmonise their national health data systems, thereby ensuring that citizens can more easily access their health data across the EU and encouraging the use of data in science and public health. The LDHC has also launched extensive work on data standardisation, in order to improve data interoperability and quality, by introducing international health data standards such as *HL7 FHIR* and *SNOMED CT*. This provides a foundation for structured, comparable, and machine-readable data.

As of 1 April 2025, hospitals in Latvia are required to enter patient discharge summaries into the national e-Health system as part of the measures taken to support the effective use of the e-Health system. To incentivise compliance, a financial mechanism has been introduced whereby failure to meet this requirement may result in a reduction of up to 10% in state funding. While the government has provided financial support for digitalisation, the uptake of digital record entry has not been uniform across all healthcare providers. As of March 2025, Latvian residents' core health data is also available to healthcare providers in the Netherlands and Malta, in addition to existing cross-border exchanges with Estonia, France, Portugal, and Spain. This

[8] The regulation governing the EHDS entered into force on 26 March 2025. The regulation distinguishes between primary data use (strengthening patient rights and system interoperability) and secondary use (for research and public health purposes).

cross-border health data exchange ensures faster, higher-quality treatment and better access to health information, especially in critical situations.

In addition to EHDS integration processes, Latvia is also involved in the "1 + Million Genomes" initiative. Currently comprising 25 EU Member States, as well as the United Kingdom and Norway, forming a network of 27 participating countries. The initiative aims to establish a cohort of at least one million sequenced genomes, thereby providing secure access to genetic and phenotypic data for research and innovation purposes.

3.4 Biobank Draft Law (Biobankas Likumprojkets)

The Biobank draft law, developed in Latvia and currently under review by the *Saeima* (Parliament), could represent a significant milestone for Latvia's healthcare and research sectors. The law aims to establish clear conditions for the creation of biobanks, the collection, storage, and use of biological samples, while safeguarding donor rights and ensuring data security. Biobanks will enable more precise and personalised treatment that is tailored to each patient's individual biological characteristics. The new legislation will not only strengthen quality and ethical standards in research, but will also foster public trust, engagement, and participation in scientific processes. This means that patients will gain access to more advanced diagnostic and treatment options in day-to-day medical practice. Overall, the Biobank Law is a long-awaited legal instrument that clarifies the legal framework, supports innovation in healthcare, and reinforces Latvia's role as a reliable research partner within the European area. It marks a step towards healthcare that is individually tailored and grounded in data and science.

3.5 Conclusions and Future Perspectives in Health Data Management

Latvia is currently undergoing significant transformations in the enhancement of its health data system by moving from fragmented data collection to an integrated, people-centred, and secure approach. Structured and digitised data, genomic data, and a new legislative framework form the foundation of a modern healthcare ecosystem.

To fully achieve this set of objectives, it is essential to improve public awareness and to ensure the interoperability of health data systems. The attitudes of both the population and healthcare professionals alike demonstrate readiness for change and a demand for effective and purposeful use of health data. Latvia has the opportunity to become a leading example in modern, patient-oriented health data management through the newly established Latvian Digital Health Centre, a clear regulatory framework, and active participation in European Union initiatives.

4 The Usage of Administrative Data in Constructing Historical Memory and Shaping National Identity

June 14 and March 25 are days of remembrance in Latvia—as well as for two other Baltic States, Lithuania and Estonia—to commemorate the victims of the forced removals of civilians, carried out by the Soviet regime in June 1941 and March 1949, during the USSR occupation of the three Baltic States.[9] In Latvia, these forced removals of civilians are widely referred to as the "deportations" or "the Soviet deportations". In total, more than 57,000 people—including approximately 15,000 minors—were arrested in Latvia, put on the cattle trains and sent away to Siberia and other remote areas of the USSR (Geisler, 2009). According to the historical data, 8 in 10 deported people were Latvians, 1 in 10 was a Jewish, and the rest represented a wide array of other nationalities, including Russians (Zālīte & Eglīte, 2001).[10] Today, the deportations form a significant part of the collective historical memory of the nation: they are discussed in school textbooks; the questions related to them are included in the history and culture tests that applicants for Latvian citizenship (through the naturalisation process) must pass; national flags must be raised in a mourning presentation and memorial services are organised across the country each year on both 14 June and 25 March.

In 2021, on the eve of the March memorial events, a private enterprise—which also happens to be one of the main service providers in the Baltic States in the field of cartography, geographic information systems (GIS), and geospatial intelligence solutions—published a commemorative interactive map that displays the names of the people deported in 1941 and 1949 from the territory of Latvia as well as showing the exact location of their place of residence at the time of arrest on the map.[11] The map was mainly created on the basis of historical data collected by the State Archive of Latvia.[12] More precisely, commemorative map developers used:

(a) the historical deportation lists from the State Archive of Latvia.[13] The lists are the official Soviet documents with the names of individuals and families deported from Latvia. The historical lists also recorded the age of individuals, their family relations, and, on some occasions, their occupation and social status;
(b) The historical address registries (last known address before the deportations took place);
(c) The historical maps to geocode the address data. Geographical distance approximation was used in cases where the historical address was found to longer exist (due to war damage, renaming, etc.).

[9] Such deportations during/after the World War II took place also in the territories of Finland, Germany, Moldova, Poland, Romania, and Ukraine as well as several countries in Caucasus and Central Asia.

[10] The data concern exclusively the deportation of 14 June 1941.

[11] The commemorative interactive map was entitled "deportetie.kartes.lv" (translation in English: "deportees.maps.lv"; see: https://deportetie.kartes.lv/lv/c___56.949466-24.103452-8).

[12] Website (in English): https://www.arhivi.gov.lv/en/department/latvian-state-archive.

[13] Available here: http://eresursi.arhivi.gov.lv/content.aspx?id=319&mainId=271.

The above-listed historical data were cross-checked and validated by the archive specialists and historians.

As a result, the functionalities of the map allow searching by both the name and geographical location, hence visualising the scale and geographic spread of the Soviet deportations from Latvia. At the same time, the map does not show all of the deported individuals' personal details, thereby reducing the information displayed to the individual's name and age. The map can be used by general public, by researchers, and by students—basically, anyone interested in finding more information about these historical events. The map is perceived by the Latvian society both as a digital educational tool and as a way of commemorating those affected by the Soviet deportations.[14] However, the underlying microdata also has the potential to be the root of a relational database where further administrative data sources (spouses and other family ties, social security contributions, health records) can be linked together to study, for example, the long-lasting effects of involuntary emigrations on family members left-behind, even across generations.

5 Conclusions

This chapter has explored the use of administrative data in Latvia through the lenses of three well-defined case studies, namely census data, healthcare information, and deportees. An overall picture of the increasing use and integration of administrative data sources has become clear, where legislative initiatives are advancing digitalisation, interoperability, and data safety. The Latvian experience also illustrates that the access to and use of administrative data cannot be separated from broader political and institutional dynamics. First, **EU conditionality has played a decisive role**, by providing both incentives and benchmarks that have shaped Latvia's trajectory. Europeanisation has not only ensured compliance with EU data standards, but has also encouraged innovation in both census methodologies and health data governance. Second, the transition towards a sustainable, high-quality administrative data system is a **long-term undertaking**. Projects such as improving data interoperability, strengthening institutional capacities, and developing robust quality assurance frameworks require consistent investment over many years. Their impact is gradual, rather than immediate, underscoring the need for strategic patience and continuity. Finally, successful reforms depend on **cooperation across a wide range of actors**—from central authorities and local governments to private providers and civil society.

[14] The contribution of the map in fostering the collective sense of belonging can be implied indirectly from the fact that it is not unusual for the website to be temporarily disrupted due to the high demand around the respective days of remembrance: see TVNET/LETA, Digitālā kartē apskatāmi 1941. un 1949. gadā deportēto Latvijas iedzīvotāju saraksti, 14.06.2025: https://www.tvnet.lv/7739221/digitala-karte-apskatami-1941-un-1949-gada-deporteto-latvijas-iedzivotaju-saraksti%3Futm_source%3Dchatgpt.com.

References

Geisler, I. E. (2009). From the voices of the deported: The mass Latvian deportation of June 14, 1941. *Oral History Forum D'histoire Orale, 29*, 1–10.

Zālīte, Indulis, Eglīte, Sindija, on behalf of the Constitution Protection Bureau's Centre for the Documentation of the Consequences of Totalitarianism. (2001). The structural analysis of the deportation of June 14, 1941. *Latvijas Vēstnesis*, June 15, 2021, No. 93.

Ieva Bloma is a member of the Strategic Analysis Team of the European Union Agency for Asylum (EUAA), where she carries out data analysis and research on asylum-related migration in the EU. In addition, she manages the EUAA foresight activities. She holds a Master in European Studies from the University of Latvia, a Master in Management of the Public Administration from Aix-Marseille University, a Master in Regulation from the London School of Economics and Political Science, and a Master of Research from the European University Institute, where she also served as research assistant.

Jānis Glazkovs is Counsellor for Digital Affairs and Regional Development at the Permanent Representation of the Republic of Latvia to the European Union in Brussels. He represents and advocates Latvia's interests in EU digital and regional policy domains, contributing to negotiations, policy coordination, and the preparation of national positions. His work covers key EU digital initiatives, including data, artificial intelligence, digital identity, interoperability and international digital partnerships, and involves close cooperation with EU institutions, international organisations, and Member State delegations.

Fabio Berton is Policy Analyst on impact evaluation at the Competence Centre for Microeconomic Evaluation at the Joint Research Centre of the European Commission. He is also Associate Professor of Economic Policy at the University of Torino, and Research Fellow at CIRET, IZA and LABORatorio R. Revelli. His research focuses on applied labour economics, and on the evaluation of labour policies in particular. Fabio has publications on The Economic Journal, The British Journal of Industrial Relations, The Oxford Bulletin of Economics and Statistics, The Review of Financial Studies, and The Journal of Economic Inequality among the others.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 18
Open Administrative Data in Lithuania

Egidijus Barcevičius

Abstract The public sector in Lithuania generates substantial administrative data; this data has increasingly been made available for use by businesses, think tanks and research institutions. Lithuania has made significant progress in opening up administrative data, as evidenced by its improving ranking in the OECD OURdata Index and the Open Data Maturity Index. The main portal for the publication of open datasets is data.gov.lt, while researchers may also request customised datasets from the State Data Agency (SDA). However, many administrative datasets still remain unopened. The opened datasets vary in size and granularity, the documentation describing the data is not always sufficiently detailed and the quality of the data itself is sometimes problematic. The use of open data has been relatively limited to date. However, recent investments and procedural development are likely to improve both access to and quality of open data, thereby leading to the emergence of new policy-relevant research.

1 Introduction

The public sector in Lithuania generates a wealth of data that is used to implement administrative and regulatory functions and to provide public services.[1] This data is increasingly being opened to and used by businesses as well as think tanks, consultancies and research institutions.

Lithuania has made noticeable progress in the last several years by opening its administrative data for reuse. According to the OECD OURdata Index, Lithuania was among the best-performing countries in 2023. The Index assesses countries based on the pillars of data availability, data accessibility and government support for data reuse (OECD, 2023). According to the Open Data in Europe report, Lithuania came

[1] I am grateful to several interviewees at the State Data Agency for sharing their views on the current status and the future of administrative data access in Lithuania.

E. Barcevičius (✉)
PPMI Group, Vilnius, Lithuania
e-mail: egidijus@ppmi.lt

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_18

7th in 2023 among the 35 participating countries[2] in terms of open data maturity. The index looks into the open data policy, open data impact, open data portal and open data quality (Data.europa.eu, 2023).

An important factor that allowed Lithuania to progress in opening the administrative data was the implementation of two projects, both co-funded with EU support. The first project developed the open data portal data.gov.lt, including the procedures for opening data and the technical infrastructure. The project was managed by the Information Society Development Committee[3] and implemented in 2016–2023. The second project, under implementation until mid-2026, develops the "data lake" by the State Data Agency and includes inventorying and opening data from multiple public sector organisations. During this project, the SDA proactively reaches out to public sector organisations, identifies the administrative datasets, provides support documenting the datasets, defines the meta-data and helps to submit the datasets to the open data portal. The SDA also ensures that data is accessible through the "data lake" infrastructure, where the data is further cleaned and verified. This approach facilitates the connection of different data sources, thereby ensuring data interoperability and reuse.

2 Avenues for Data Access

Lithuanian public sector institutions differentiate between (1) published data; (2) open government data and (3) data which can be accessed upon request.

The first category concerns data that institutions choose to publish on, for example, their websites in order to inform the wider public about their activities. It is of limited use for public policy research, but may still be beneficial as a source of contextual data.

Next, open data means the raw data, which is depersonalised and available for reuse, free of charge, in open and machine-readable formats. In Lithuania, the main open data portal is data.gov.lt, although individual public sector organisations and municipalities can also publish their data on their own platforms.[4]

Finally, some data collected by public sector institutions cannot be readily published online due to technical, procedural or data protection reasons. If researchers need a customised dataset, which is not available through the open data portals and which combines data from different public sector information systems and registers, then they may reach out the State Data Agency. The Agency will process the request and will provide data if the request is approved.

[2] EU-27, plus EFTA and EU candidate countries.

[3] This is an agency under the Ministry of Economy and Innovation; it was renamed into State Digital Solutions Agency since August 2024.

[4] For example, open data by Vilnius municipality can be accessed at open.vilnius.lt.

3 Current Status of Data Access

At the time of writing in late July 2024, 184 public sector organisations have registered in the open data portal data.gov.lt as providers of open data and the portal provides access to 1,580 datasets. This still represents a relatively small number of all of Lithuania's organisations and data sets (Information Society Development Committee, 2023). However, the organisations that have already opened some of their data include the major ones with the most extensive and comprehensive datasets. These include the Center of Registers, the State Social Insurance Fund (Sodra), the State Tax Inspectorate (VMI), the State Data Agency and others. Furthermore, the SDA reports that they are currently in touch with 362 public sector organisations, have identified 1461 data sources and have already opened 163 datasets for the "data lake" project.[5]

The open data portal data.gov.lt is organised into 14 thematic categories, covering areas such as environment, economy and finance, energy, culture and others. Individual datasets can be downloaded in machine-readable formats (.CSV,.JSON,.RDF,.TXT) and they can be accessed through an API. Examples of the datasets include:

- Environment: hydrological observations, data on building energy performance certificates;
- Economics and finance: support for small and medium-sized businesses; procurement, amount of taxes paid by legal entity;
- Social security: refugees arriving at the Refugee Reception Centre, data on social support provided to students, child benefits data;
- Education: workload for educators, non-formal education providers for children;
- Economic activity and business: Real estate registry (NTR) apartment purchase transactions, producer price indexes.

However, the open data portal data.gov.lt has limitations. It is organised as a collection of separate datasets (each associated with a specific institution) that vary in size and significance. It is difficult or outright impossible to combine different data sets. The granularity of data may differ, the documentation concerning specific datasets and variables is uneven and the datasets may contain missing values. Variables containing personal data, which could be used to link different datasets, obviously cannot be part of datasets on the open data portal. Furthermore, the data.gov.lt has very limited functionalities for rendering geospatial data on maps and for exporting such data. Therefore, such data is provided through other portals[6] and this fragments access to geospatial data.

Importantly, the situation concerning open data has been changing dynamically, due to investments being made in technology as well as procedural changes. New administrative datasets are being published regularly on the open data portal.

[5] 7 July, 2024. Dashboard for monitoring the "Data lake" project: https://experience.arcgis.com/experience/03398d383eca47e4a17cc853d72df2d4.

[6] https://open-data-ls-osp-sdg.hub.arcgis.com; www.geoportal.lt.

The portal also has a data request function; the requests are being monitored and responded to by the State Data Agency.

If data is not readily available on the open data portal, then researchers may reach out to the SDA and can request access to a customised dataset. Importantly, the SDA may combine different datasets at the level of an individual person by removing personal identifiers (e.g. names, surnames and identity codes) and by creating unique codes that could be used to link datasets related to one specific request.

Researchers may request access to data by filling out a form[7] that asks them to list the variables of interest to their research. After receiving an application, the SDA then assesses its objectives and examines whether the data exists and can be provided without compromising personal data protection. If the SDA decides to grant data access, then it also indicates the time needed to prepare the data and the related fee. The time to prepare the data and the fee depend upon the complexity of the request, for example, if data is already available within the "data lake" and whether or not the SDA needs to obtain data from other data controllers. The fee is intended to compensate for the work of the SDA's data scientists and programmers; there is also a charge that is applicable specifically for health data; data controllers can often also charge an additional fee. The time required to fulfil the data request may range from a week to months, whereas the fee may start at several hundred EUR and reach several thousand EUR.

The SDA fulfils the request by creating a customised dataset. It provides access to this dataset through a "sandbox" on the SDA's "data lake" or it can also choose to publish the dataset on the open data portal for public access.

As of the time of writing, the number of requests submitted to the SDA for accessing data for research purposes has been relatively low. This is because the SDA's relevant technical and human resource capacities have only been developed recently. In total, about 10 requests for health data and several requests for other types of data were submitted in 2023. Consequently, the SDA has not yet encountered enough use cases to develop a comprehensive set of rules that are applicable to various situations. Procedural practices for evaluating data requests, including pricing and the assessment of the goals of data access requests are being discussed frequently. Additionally, the portals providing access to open data (including navigation and access forms) primarily function in Lithuanian, which limits access for researchers located in other EU or non-EU countries.

4 Thematic Overview

Labour Market, Social Situation, Population Data. The key institutions that collect the data are the State Data Agency, State Tax Inspectorate (tax data), State Social Insurance Fund (Sodra), Employment Agency (registered unemployment) and the Ministry of Social Affairs and Labour (data on social payments and other types of

[7] The form can be accessed at: https://osp.stat.gov.lt/e.-paslaugos.

support for families). The data can be used, for example, to understand the effects of social support or unemployment support on employment and income. In 2023, analysts from the Ministry of Social Affairs and Labour used data combining administrative data sources on the debts of individuals, income and employment, which showed that around 70% of individuals who have overdue debts and are subject to bailiff proceedings do not have any formal employment (SADM, 2023a). This shows that individuals with overdue debts are being pushed towards inactivity in employment or informal employment. In response, the Ministry initiated and the Parliament approved a change in the law, which introduced "debtor holidays" allowing for a pause in debt collection for people in employment (SADM, 2023b).

Health Data. Lithuania collects a wealth of health data about individuals through the e-health system, including health checks, health conditions and treatments. The Health Insurance Fund (VLK) collects data about the revenues and expenses at the various levels of the health system, including expenses for medicines, prescriptions and treatments. Other data is collected by individual agencies, centres and laboratories under the Ministry of Health as well as by individual health services providers. The data has been used by researchers to assess the effectiveness of various treatments, such as to monitor the proliferation of COVID-19 and the effects of vaccines against the virus.[8]

Business Company Data. The data is collected in a variety of registers, such as the register of legal entities, the information system concerning participants of legal entities (JADIS), State Tax Inspectorate (VMI) (tax and salaries' data), Bailiff Information System (AIS), Real Estate Registry, EU grant registry, Agricultural Registry, GEO-information data and others. The data is used by businesses, including banks, insurance and fintech companies. Some businesses have built their business models by relying on access to administrative data, such as by combining and selling reports on companies, credit ratings.[9] In 2023, researchers and activists used data made available by the State Data Agency, which combined data from the Agricultural Registry, State Social Insurance Fund (Sodra), Civil Registry, EU Grant Registry, to analyse the impact of farms growing fur animals on employment. The research found that a limited number of fur farms accounted for around 90% of all fur animals in Lithuania, that the number of persons working in this business is small and that those employed in this business have other employment opportunities in their regions. The research played a role in the decision by Lithuanian parliament to ban fur farms in Lithuania (LRT, 2023).

[8] For example, open data on Covid-19 incidences and vaccination can be accessed at https://opendata-ls-osp-sdg.hub.arcgis.com/pages/84bdca2fdcb8491ea36828555e81d18f.

[9] For example, rekvizitai.lt, creditinfo.com, carvertical.com, trafi.com.

5 Conclusion

Access to administrative data for researchers has improved over the last several years. The open datasets are available through the open data portal data.gov.lt, with new datasets being added to this portal regularly. However, many datasets still remain to be opened and the open datasets vary in terms of their size and granularity, the documentation of datasets is not always sufficiently detailed and access to geospatial data is fragmented. It is also difficult or even outright impossible to combine individual datasets.

Researchers have an opportunity to reach out to the SDA and request a customised dataset that may combine data from different sources for more complex requests. However, this option has not been commonly used yet because the SDA has only recently gained the necessary technical and human resource capacities and some researchers also find the fee to be too high. With a limited number of use cases, the pertinent procedural rules and practices are still under development. Significant advancements in evidence-based research with policy implications can be expected as researchers experiment with the new possibilities for both accessing and combining open government data.

References

Data.europa.eu. (2023). *2023 Open data maturity report*. Publications Office of the European Union.

Information Society Development Committee. (2023). *Report on opening public sector data* (in Lithuanian), available online. at data.gov.lt.

LRT. (2023). *After intense discussions, the Seimas decided: The fur farming industry will be banned* (in Lithuanian). https://www.lrt.lt/naujienos/verslas/4/2082127/po-inirtingu-diskusiju-seimas-apsisprende-kailiniu-zvereliu-verslas-bus-uzdraustas

OECD. (2023). *2023 OECD Open, Useful and Re-usable data (OURdata) Index: Results and key findings* (OECD Public Governance Policy Papers, No. 43). OECD Publishing.

SADM. (2023a). *4 out of 5 people in debt are not in employment* (in Lithuanian). https://socmin.lrv.lt/lt/naujienos/4-is-5-darbingo-amziaus-skolininku-nedirba-sadm-siulymas-diferencijuotas-skolu-isieskojimas/

SADM. (2023b). *Incentives for debtors to work legally and escape the debt trap* (in Lithuanian). https://socmin.lrv.lt/lt/naujienos/skolininkams-paskatos-dirbti-legaliai-ir-isbristi-is-skolu-liuno-o-dirbantiems-seselyje-grieztesne-kontrole/

Egidijus Barcevičius is Research Director at the PPMI Group in Vilnius. He has more than 20 years' experience in conducting policy-relevant studies and evaluations and providing advice to decision makers. His main field of expertise is the digitalisation of public administration, industries and labour markets. He has a strong track record of working with and leading multinational teams of experts. He has worked for a variety of national and international clients at various levels of government: municipal authorities, national ministries, the European Commission and EU agencies. He conducts training on public policy analysis and is author of numerous publications.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 19
Administrative Data Access in Luxembourg



Michela Bia

Abstract Access to administrative data in Luxembourg is both guaranteed and facilitated through the Luxembourg Microdata Platform (LMPD). This centralised platform serves as a valuable resource for both researchers and policymakers by offering streamlined access to a wide range of administrative datasets from various government agencies. The LMPD simplifies the data acquisition process by consolidating these datasets, thereby saving researchers time and resources. Its user-friendly interface and robust security measures ensure that sensitive data remains both protected and accessible for legitimate research purposes. Consequently, policymakers can leverage the platform's wealth of information for policy evaluation studies. There are, however, areas in which improvements are possible, beneficial, or even critical, such as data quality, dataset access (e.g. to census data), and confidentiality laws. Luxembourg could further harness administrative data's potential by adopting more flexible and collaborative data governance approaches in order to effectively address complex societal challenges, advance research and inform comprehensive evidence-based policymaking.

1 Overview of the Chapter

This chapter explores the opportunities and challenges associated with accessing administrative data for research purposes in Luxembourg, with a focus on the role played by the IGSS (Inspection Générale de la Sécurité Sociale) microdata platform. This platform is primarily intended for researchers and policymakers who rely on administrative data to inform their studies and evaluations. It also aims to provide them with a comprehensive understanding of the mechanisms that either facilitate or hinder access to valuable administrative datasets.

This chapter highlights the positive strides that Luxembourg has made in centralising administrative data access through the IGSS microdata platform. This platform represents a significant advancement in supporting policy evaluation efforts

M. Bia (✉)
LISER, Esch-Sur-Alzette, Luxembourg
e-mail: michela.bia@liser.lu

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_19

by providing streamlined access to a variety of datasets from different government agencies. Its practical aspects are also evaluated, with particular attention to ease of use, security features, and the efficiency it offers to researchers seeking access to administrative data. Conversely, it identifies both the challenges and limitations that still exist, such as issues surrounding access to certain sensitive datasets (e.g., census data) and confidentiality laws. These barriers can impede research efforts and can serve to limit the scope of policy evaluation. Finally, this work recommends improvements be made in order to enhance the quality and usability of the data and to encourage the adoption of more flexible, collaborative approaches to data governance that can further unlock the potential of administrative data for innovative research and broad evidence-based policymaking.

In terms of the legislative framework, while the IGSS microdata platform simplifies data access, it operates within Luxembourg's broader legal framework for data protection and confidentiality. For this reason, the relevant laws that govern data access and confidentiality, which ensure compliance with the General Data Protection Regulation (GDPR), are also presented and described briefly. Understanding these regulations is critical for researchers who must balance the need for data access with legal and ethical obligations.

The remainder of this chapter is organised as follows: Sect. 2 introduces the Luxembourg Microdata Platform (the IGSS), the related Data Dictionary, the so-called proportionality rule that is tailored to suit each individual data request and how IGSS data can be merged with external data. Section 3 describes the procedure and the ASK4MDP application, which eligible users must use to submit IGSS data access requests via the Microdata Platform. Sections 4–8 offer descriptions of various policy evaluation studies that were developed using local administrative datasets and provide an overview of the current state of administrative data usage and access. The focus will be on labour markets, education, family, health and housing market analyses, respectively. Section 9 concludes.

2 Introduction to the Luxembourg Microdata Platform

In Luxembourg, the Microdata Platform on Labour and Social Protection (LMPD) is the primary means of accessing personal data intended for research endeavours that focus on policy evaluation. These datasets consist of administrative data that are centrally managed by the '*Inspection générale de la sécurité sociale*' (IGSS) and which, as per its responsibilities, is predominantly collected by the country's social security institutions. The available information in the social security-based datasets extends beyond the norm, as discussed later in this chapter.

The growing need for highly detailed research data highlights the critical importance of establishing technical and organisational protocols to safeguard the confidentiality of data, as highlighted by the IGSS itself in its methodological note[1] (Thomas & Zanardelli, 2023). To that end, the IGSS partnered with the Ministry of Labour, Employment and the Social and Solidarity Economy to create the LMPD in 2018 in order to tackle the challenge of balancing data protection with accessibility. Since its launch, the platform has undergone significant development, drawing upon insights gleaned from its involvement in diverse projects.

The protocols and safeguards instituted by LMPD have undergone continual refinement and adaptation in order to be aligned with the evolving demands of research, data protection norms and regulatory mandates. The LMPD, which initially centred on cash benefits within Luxembourg's social security framework, has broadened its purview to encompass non-monetary benefits. This expansion commenced during the COVID-19 pandemic, during which time the platform played a pivotal role in disseminating pandemic-related data. The platform's protocols and data protection measures have also been fine-tuned, thereby ensuring compliance with the General Data Protection Regulation (GDPR). The LMPD adopts a holistic approach in order to effectively balance the dual objectives of preserving data integrity and facilitating research accessibility (Thomas & Zanardelli, 2023). The following information is provided within the IGSS methodological note:

- Application of eligibility criteria to requests for accessing data;
- Deployment of a data dictionary to automatically safeguard data accessible for research;
- Conducting proportionality assessments customized to each data request, prioritizing the need-to-know principle and employing tailored protective measures;
- Peer review of privacy risks linked to data requests and implemented protective measures;
- Signing of a confidentiality agreement outlining the user's obligations;
- Implementation of a systematic procedure for validating outputs.

These guiding principles and protective measures concerning personal data closely resonate with the objectives outlined in the EU—Data Governance Act. These measures are geared towards establishing a comprehensive framework that instils trust and confidence in data sharing practices.

2.1 The Data Dictionary

The IGSS provides a detailed description of the data available to support research studies, compiled in the so-called Data Dictionary[2] (DD), which is available in both

[1] https://igss.gouvernement.lu/fr/publications/apercus-et-cahiers/cahiers-methodologiques/202305.html.

[2] https://igss.gouvernement.lu/dam-assets/microdata-platform/data-dictionary.pdf.

the English and French. The DD contains the following information: individual sociodemographic characteristics, child–parent relationships, spousal relationships, household characteristics, labour market data, family benefits, social assistance benefits, and health data, among others.

As reported by Thomas and Zanardelli (2023), ensuring robust data protection through the data dictionary's privacy-centric design is one of the keystones of the LMPD. While the DD does not include direct identification variables, such as names or addresses, some variables may indirectly facilitate identification, either through specificity or in combination with other variables. The DD's development was based on two protection levels in order to address this issue:

- Variables marked with the symbol 🔒 are aggregated by default in order to mitigate re-identification risks while maintaining analytical value. For instance, nationality data, though available in detail, is aggregated so as to safeguard against re-identification risks. IGSS has pinpointed specific nationalities or demographic cohorts that are crucial to understand country-specific attributes, thereby ensuring ample sample sizes to minimise identification risks.
- The second tier of data protection is denoted by the symbol 🔒🔒: variables marked with this symbol were not initially provided to the researcher. Due to their sensitivity, this information is safeguarded to prevent the disclosure of sensitive data in the event of re-identification. This protection extends to data such as minimum income receipts or work absences. However, these variables may be made available for research purposes after the removal of identifiers, provided that alternative protective measures are implemented.

2.2 The Proportionality Rule

The principle of proportionality is applied on a case-by-case basis to each data request at IGSS. Each request is allocated to a designated expert within IGSS, such as someone who oversees the entire process, including conducting the proportionality analysis. Requests are assigned to IGSS experts according to their area of expertise. The application of the need-to-know principle involves several aspects (see Thomas and Zanardelli (2023) for further details on the procedure):

- Project Relevance[3]: researchers outline their project's objectives in their request. If the analyses can be conducted with aggregated statistics, instead of individual-level data, then experts provide the statistics accordingly;
- Scope of Study: researchers specify the population under study, such as all Luxembourg residents or only the active workforce. The proportionality analysis ensures that the data provided aligns with the research question;
- Timeframe: researchers indicate the desired data period. If it is deemed to be excessive, then IGSS experts may shorten it;

[3] See Sect. 2 for further details about eligibility rules.

- Variables Selection: researchers choose variables from the data dictionary, providing justifications where there are deviations from the default level of granularity. Each variable undergoes a proportionality analysis independently.

IGSS's approach to data protection is comprehensive and applies universally to all data types, including health data, which is known for its heightened sensitivity compared to, for instance, employment data.

2.3 LMPD Data Merged with External Datasets

The database is structured into thematic registers (currently totalling nine) in order to encourage a 'Data for Research approach' at IGSS, and covers diverse topics such as: individual sociodemographic characteristics, nature of affiliation to the social security system, employment details, firm-level characteristics, social benefits and minimum income, absenteeism at the workplace, professional relocation, household attributes and family links. These registers can be interconnected by using pseudonymised identifiers, thereby facilitating flexible data analysis as per researchers' methodologies. Data is available monthly from 2002 onwards, thereby enabling in-depth analysis of long-term trajectories (where needed) for research inquiries.

Linking LMPD data with other databases within the Luxembourg statistical system is also facilitated and this allows for the integration of supplementary information. For instance, data from the Luxembourg employment development agency (ADEM) can be linked to the LMPD dataset and this broadens research possibilities (e.g., the evaluation of active labour market programmes for the unemployed). Stringent protection measures, including rigorous pseudonymisation procedures, are in place in order to safeguard data and individual privacy. Data linkage is often utilised to augment surveys based on LMPD data and this enhances research outcomes. However, it is important to highlight that explicit consent from survey participants is necessary for data-survey linkage; this ensures transparency and respect for participants' privacy preferences.

The potential to link LMPD data with several (local) external sources, and to introduce new variables (often tailored to researchers' specific requests), into the data dictionary clearly presents a valuable opportunity to conduct impact evaluation studies based on administrative data. It serves as an effective tool to provide policymakers with better-informed insights on various public policies that are being implemented in Luxembourg, such as those concerning the labour market, health, education, family or housing policies. Conversely, there are also opportunities to enhance certain variables within the IGSS platform. For instance, the categorisation of occupation currently distinguishes between individuals working solely in the public versus private sectors or as self-employed. However, details such as whether individuals are employed part-time or full-time are not provided, even though this information could potentially be inferred from the number of hours worked per month. Additionally, critical variables, such as education, are only available for a

subset of individuals, specifically the unemployed, through the merging of the IGSS dataset with the ADEM database.

Another complication in the IGSS' administrative data may arise when some individuals are 'lost' due to their departure from the country. This leads to missing values because the dataset is not updated to reflect changes in population registration. Consequently, it becomes impossible to distinguish whether these individuals are inactive or have simply left the country, given that there is no mechanism in place for deregistering those who emigrate. Furthermore, dimensions such as family and school allowances or child bonuses, crucial for evaluating the impact of important public policies like childcare interventions, are absent for certain segments of cross-border workers due to attrition issues. It is important to note that attrition is a recurring challenge in Luxembourg's administrative data, particularly due to the significant presence of cross-border workers. Consequently, there is a risk of losing long-term labour market outcomes for unemployed individuals who had been cross-border workers, as they may find employment outside of Luxembourg over time. Addressing this challenge requires linking multiple administrative data sources from neighbouring countries, such as Belgium, France and Germany, and those in Luxembourg.[4] This would enable the recovery of pertinent information, thereby facilitating comprehensive impact evaluation studies.

3 Data Access: The ASK4MDP Application[5]

Eligible users seeking data access[6] through the *LMPD* include researchers affiliated with Luxembourg-based organisations who are aiming to conduct research projects for statistical or scientific purposes.[7] This affiliation ensures that research projects receive validation and support from the organisation's management. Furthermore, the focus on statistical or scientific purposes aligns with the GDPR's authorisation for secondary data use. *Data pseudonymisation* plays a crucial role in making microdata available under the GDPR. Researchers can access valuable data while protecting individual privacy by anonymising personal information. This approach maintains a balance between data utility and privacy rights.

[4] The country has made significant investments in cutting-edge digital infrastructure, such as HPC and Meluxina ((Meluxina: Luxembourg's new supercomputer—Luxembourg (public.lu), and more recently in improving the linking and exchange of data across organisations through the Luxembourg National Data Service (LNDS, https://data.lnds.lu).

[5] For further details on the application procedure, see the '*The Life Cycle of a data request integrated in a digital solution*' included in the IGSS' Methodological note (Thomas and Zanardelli, 2023).

[6] The Data access request is possible only via an authentication mode (LuxTrust or eIDAS) through the LMPD platform (which includes the ask4mpd application form).

[7] By statistical (1) or scientific (2) purposes, the IGSS refers to whether the project focuses on statistical inference in case (1) (e.g., descriptive statistics or associations among different factors of the study under analysis) or generating new knowledge, testing new theories in case (2) (e.g., the impact of a policy with a societal implication).

The data request involves a series of critical steps and the *ask4mdp* application plays a central role in streamlining this process. The key stages of this process are outlined below:

(1) *User Account Creation*: researchers must create a user account to access *ask4mdp*. This step ensures proper authentication and access control;
(2) *Data Request Preparation*: once the account is activated, researchers can describe their data request using an online form. The form includes sections for project details, eligibility criteria and project objectives. Experts from the IGSS (General Inspectorate of Social Security) assess the request based on these details;
(3) *Variable Selection and Justification*: researchers select the specific variables that they need for their analysis. The application integrates a data dictionary (see subsection 1.1) with available variables and their values. Researchers must justify each variable based on their research question. Additional justification is required if a higher level of granularity is requested;
(4) *Submission to LMPD*: after completing the request form, researchers submit it to the LMPD; the request becomes final and cannot be modified further thereafter. Automatic notifications inform the researcher of the submission and the IGSS experts of receipt;
(5) *Allocation to an IGSS Expert*: an auto-assignment system allocates the request to an IGSS specialist. The specialist focuses on the relevant area (e.g., labour market, social protection or health). If the request spans multiple areas, then the IGSS experts decide who will handle it;
(6) *Eligibility Assessment***:** the IGSS expert assigned to the request evaluates the researcher's eligibility. Three critical questions are considered: Does the project require microdata? Is the researcher authorised to use microdata? Is the project purely statistical? All three questions must receive a positive answer to proceed. Proportionality analysis is initiated, focusing on data fields, time periods and requested variables. Justifications provided by the researcher guide the validation or invalidation of each variable and any requests to modify granularity levels. A messaging system facilitates further information exchange;
(7) *Proportionality Analysis and Variables List:* the proportionality analysis generates a list of variables based on the researcher's needs and the need-to-know principle. This list encapsulates privacy protection measures. It is attached to the confidentiality agreement, always associated with a closed list of variables. The application automates the variables list generation;
(8) *Double Validation:* first, the request processing undergoes peer review by the IGSS data team. Second, the researcher receives access to the validated request, is informed about decisions made and about the final variables list. Researchers can question decisions if they impact project feasibility. Close communication with IGSS experts minimises potential issues.

(9) *Role Allocation*: after processing validation, the operational phase begins. The expert overseeing the request defines necessary roles and assigns colleagues accordingly.
(10) *Request Implementation*: i) virtual desktop preparation (user accounts, software installation); ii) data extraction by data extractors using the variable list; iii) confidentiality agreement preparation (automatically generated based on the provided information). The application updates the request status and displays it on a dashboard. Contract dispatch occurs upon final task validation. Once granted, data access is free of charge and allowed via remote access to the LMPD by using an authentication mode (e.g., LuxTrust or eIDAS) from personal computers.

The sections that follow will offer descriptions of various policy evaluation studies that were developed utilising local administrative datasets. They will provide an overview of the current state of administrative data usage and access.

4 Labour Market

Bia et al. (2022a, 2022b) conducted a study evaluating the impact of language training programmes for unemployed individuals in Luxembourg as part of the Evalab4Lux[8] Project. They used administrative data from the country's labour force and combined two comprehensive datasets.

The first dataset comprised a matched employer–employee database that was sourced from the global social security database in Luxembourg (IGSS). This dataset included workers' trajectories from their initial entry into the labour market. The second dataset was a panel dataset on training programmes[9] that was collected by the Employment Agency (ADEM) in Luxembourg. The observation unit in this study was represented by an 'unemployment file' that corresponded to a spell of unemployment. Each individual's registration with ADEM resulted in the opening of an unemployment file; this file was closed when they no longer checked in at scheduled agency meetings (e.g., upon finding a job, missing meetings or exiting the labour market).

The collected data included background characteristics (e.g., gender, age, nationality, civil status, education, health status), labour history (e.g., job start date, wage, months worked in the last 12 months, sector of activity, previous employment status), ADEM registration details (e.g., registration date, duration in months) and treatment data (e.g., type of interventions assigned, programme start date). Additionally, the ADEM dataset featured an 'employability score.' This score is a categorical variable that assessed the employability level of each unemployed worker based on various

[8] Evaluation of Active Labour Market Programs in Luxembourg (ESF Project).

[9] The period covered spans from January 2007 to June 2016. It should be borne in mind that the LMPD IGSS platform was not available between 2015 and 2017; hence, all of the data used in the analysis have been stored and processed on Liser's premises.

pre-treatment diagnostics, such as: (1) social context; (2) employment history; (3) health condition; (4) psychological condition; (5) job prospects; (6) childcare; (7) mobility and (8) job search behaviour and assessment of the unemployed profile. This variable, which is generally not available in similar studies, provided valuable information. This variable acted as a proxy for generally unobservable subjects' characteristics, such as personality traits and expectations, and skills and motivation, strongly related to subjects' choices and behaviours.

The set of pre-treatment variables used in the analysis covered many of the standard personal characteristics that the literature concerning the evaluation of training programmes considers crucial in order to avoid biased estimates of the causal effects of active labour market policies (see, e.g., Biewen et al., 2014; Lechner & Wunsch, 2013).

Their study revealed that language training programmes effectively increased reemployment probabilities in both the short and long terms, with a significant positive effect of 1.12 Euro on hourly wages detected only in the long run. These results may assist policymakers, particularly the Luxembourgish Employment Agency (ADEM), in designing future policies that are better targeted and better tailored to achieving the intended labour market outcomes in the context of language training programmes.

5 Education

Bousselin et al. (2026) study the effects of expanding the availability of 'Early Childhood Education and care' (ECEC) slots, in place of residence, on the skills of children from low-income families in the framework of the ChilDev[10] Project. The infrastructure consolidates administrative data from various sources and covers children, their families and Early Childhood Education and Care (ECEC) providers throughout Luxembourg. These sources include the national social insurance system (IGSS) and educational registers from the Ministry of Education. Each dataset provides unique identifiers for individuals and ECEC providers, thereby facilitating exact matching between sources.

The authors gather demographic and employment data from the National Insurance System for individuals as well as information on childcare companies (size, creation date) and employees (tenure, earnings). These monthly records span from 2002 to the present and are used to assign children to treatment and control groups based on the residence municipality.

They have also merged employer-employee data in order to assess the inputs of ECEC providers, such as wages and available places, since 2002. Additionally, administrative records for the ECEC system offer details about childcare usage (type, duration and cost), for each child and provider, dating back to the introduction of the

[10] hInvestigating the consequences of public investments in Early Childhood Education and Care services for the quality of provision and child development in Luxembourg (FNR Project).

Childcare Service Voucher scheme that was introduced in Luxembourg in 2009. School achievement records provide outcomes for children in public education, measured by teachers at ages 5, 7, 9 and 11. These outcomes align with the Ministry's curriculum, defining developmental competences for each child throughout their formal education.

Access to the data is granted through secure remote servers managed by the CTIE ('*Centre des technologies de l'État*') and, thus, requires strong authentication via a LUXTRUST certificate. Each individual is assigned a pseudonymised identifier that was generated by IGSS, with the link between personal and pseudonymised identifiers being stored separately in order to ensure data anonymity.

The administrative data includes details on children such as gender, age, childcare attendance and primary school test results. For parents, this information covers location, migration background and annual earnings and encompasses market wages and public transfers.

The database developed in this project establishes an innovative research infrastructure with far-reaching implications for future studies, the research community, local authorities and the general population alike. This infrastructure is designed to be archived and expanded both during and beyond the project's duration. Potential extensions include incorporating new data about educational performance throughout the life course of individuals in the cohorts analysed, along with information about their family backgrounds. These extensions could involve merging individual-level data from additional register sources to create a new longitudinal dataset that focuses on school performance for selected cohorts.

At the time of writing, the ChilDev Project is a work in progress. Preliminary results will soon be available at the following link: https://liser.elsevierpure.com/en/projects/investigating-the-consequences-of-public-investments-in-early-chi.

6 Family

Bia et al. (2025) investigate the causal impact of parental leave uptake on the post-birth hourly wage distribution of eligible mothers in Luxembourg in the context of the PARENT[11] project.[12] They rely on the IGSS database that contains comprehensive information on all parents and their children with social security numbers in Luxembourg, along with workplace details. The analysis focuses on first-time mothers who met the parental leave eligibility criteria and delivered between 2005 and 2010 (due to data limitations preventing the linkage of parental leave information for mothers with multiple births prior to 2016). The data includes extensive labour market histories from six months before conception, demographic characteristics at birth, parental

[11] Evaluation of parental leave in Luxembourg, focus on couples' strategies and the role of workplace characteristics (FNR project).

[12] See also Valentova (2019, 2024), and Valentova et al. (2022) for additional related studies published in the context of the PARENT Project.

leave uptake dates and labour market outcomes approximately 2, 3, 4 and 5 years post-birth for each eligible mother, with some variables derived from records of the National Health Fund. The sample is further restricted to include observations with labour market histories for the newborn's father, given that paternal characteristics are crucial predictors of parental leave participation and maternal labour market outcomes; this is something that is often overlooked in studies due to data constraints. Access to the data was facilitated through secure remote servers that were managed by the CTIE and that required robust authentication via a LUXTRUST certificate. Findings from estimated bounds on quantile treatment effects reveal both substantial and statistically significant negative impacts (i.e., wage reduction) on most quantiles above the median of the wage distribution (e.g., in the order of 10–15%). The estimated results suggest that the labour market interruption, caused by taking parental leave, may have significant implications for the gender wage gap. The authors primarily attribute the estimated wage penalties to the lack of accumulation and depreciation of human capital resulting from intermittent labour market participation during parental leave.

7 Health

The APPEAL[13] project represents an important research initiative. One notable study within this project, conducted by Moran et al. (2024) in the realm of health policy evaluation in Luxembourg, seeks to assess the efficacy of primary care services in Luxembourg. The focus is on the impact assessment of the Referring Doctor (RD) policy; this policy was introduced in the country in 2012. This policy encourages patients to enrol with a primary care provider to ensure coordinated healthcare delivery. The authors accessed individual-level pseudonymised data sourced from the Luxembourg National Health Insurance Fund, covering the timeframe from 2010 to 2018, for the analysis. The study primarily targets individuals diagnosed with type 2 diabetes, a medical condition that is well-documented in terms of its potential for effective management within primary care settings.

Moran et al. (2024) employed a methodology where patients' type 2 diabetes diagnoses were inferred from data on reimbursed prescribed medications, given the absence of comprehensive diagnostic records in the database. This approach successfully identified 16,775 individuals who were prescribed oral diabetes medications between 2010 and 2012. The sample was further expanded through a sensitivity analysis to include an additional 6,769 patients who received both oral medication and insulin.

The results unveiled the RD programme's limited impact on various care quality metrics and the use of prescribed medications among individuals with type 2 diabetes. Moving forward, researchers may consider broadening their scope to encompass wider patient cohorts and leveraging the data linkage with IGSS in order to gain

[13] Assessing Primary Care Performance in Luxembourg (FNR Project).

deeper insights into patients' clinical outcomes and socio-economic backgrounds. The paper is available at the following link: https://liser.elsevierpure.com/en/persons/val%C3%A9rie-moran/publications/.

8 Housing Market

The primary aim of the CRHOUSINQ[14] is to delve into the dynamics of Luxembourg's economic growth and spatial disparities over the last twenty years. Docquier et al. (2024) seek to uncover the underlying drivers of inequality and their implications for long-term sustainability by closely examining factors, such as labour mobility patterns, the concentration of economic activities across different regions, core-periphery divides and the evolution of the housing market.

To achieve this, the authors leverage comprehensive data from the IGSS that provide detailed information about income distribution, including various forms of transfers. By merging this data with real estate ownership records, obtained from the RPF ('*registre de la propriété foncière*') database, the authors have been able to construct a rich dataset covering the entire population and the real estate market landscape since 2002.

Their analysis involves categorising real estate assets by type, including land parcels, residential properties (houses, apartments) and commercial properties. They can generate insights into the overall wealth distribution and inequality dynamics over time by assigning monetary values to these assets.

This holistic approach provides a comprehensive picture of the sources and evolution of inequality in Luxembourg. The objective is to provide valuable insights that can inform policy decisions aimed at promoting more inclusive and sustainable growth in the region by understanding the interplay between economic factors, spatial dynamics and housing market trends.

However, as highlighted by Frédéric Docquier (the project's PI) during a presentation, this study is currently facing some issues. The Luxembourg Census stands out as the primary repository of comprehensive data detailing education levels, employment sectors and occupational profiles of resident workers. STATEC, the national statistical institute, employs respondent IDs (numéro national) to link census data with other sources like IGSS, RNPP and RPF databases, thereby aiming to minimise data gaps. Once this consolidation process concludes, STATEC removes the respondent IDs, preventing future linkage of census data with external sources. Given the substantial investment in census data collection, there is the potential to revise legislation to enable researchers to leverage census data more effectively.

Furthermore, STATEC's organic law (art. 16, 4) prohibits the sharing of confidential data with external researchers and users, even if stringent pseudonymisation

[14] Cross-border mobility, housing market development and spatial inequality in the Greater Region (FNR Project): https://liser.elsevierpure.com/en/projects/cross-border-mobility-housing-market-developments-and-inequalitie.

protocols are implemented by public agencies such as IGSS or LNDS. Despite the significant costs associated with data collection, many STATEC databases remain underused, thereby limiting their potential for thorough analysis and the formulation of effective policies. Addressing these protective measures is crucial to fostering research activities and facilitating evidence-based policymaking.

9 Conclusions

By restricting access to confidential data and preventing the linkage of different datasets, the current legal framework limits the scope for robust analyses and evidence-based policymaking. This limitation inhibits researchers from fully utilising available data resources, thereby hindering their ability to address complex societal issues comprehensively.

Some reforms may first be necessary to address these challenges and to unlock the full potential of administrative data for research and policy purposes. First, revisiting the Luxembourg laws governing data access and confidentiality, within the boundaries of the GDPR, could enable researchers to utilise census data more effectively while still safeguarding privacy through appropriate pseudonymisation measures. In particular, we could foster collaboration and promote rigorous analysis by revising STATEC's organic law in order to allow for the sharing of confidential data with external researchers (under strict pseudonymisation protocols). Such interventions would not only facilitate academic research, but would also empower policymakers with better evidence for decision-making.

Overall, access to administrative data is not only guaranteed, but also facilitated, through the microdata platform at IGSS and this represents a significant step forward for policy evaluation efforts in Luxembourg. This centralised platform serves as a valuable resource for researchers and policymakers alike, thereby providing streamlined access to a wide array of administrative datasets from various government agencies. The IGSS microdata platform simplifies the data acquisition process by saving researchers valuable time and resources and by consolidating these datasets in one accessible location. Moreover, the platform's user-friendly interface and robust security measures ensure that sensitive data remains protected while still being accessible for legitimate research purposes. As a result, policymakers can leverage the wealth of information available through the IGSS microdata platform in order to conduct policy evaluation studies. However, addressing the issues surrounding dataset access (e.g., to Census data) and confidentiality laws is crucial. In addition, improving specific variables (such as occupation or education within the IGSS platform), regularly updating the dataset with information from the population register and increasing the availability of longitudinal data would substantially strengthen the quality and usefulness of the data. These improvements would enable researchers to better track changes over time, gain deeper insights into long-term trends and policy effectiveness, and foster innovation, advance research and support more comprehensive evidence-based

policymaking. By adopting more flexible and collaborative approaches to data governance, Luxembourg can harness the full potential of administrative data to inform solutions to social challenges, thereby providing a richer and more comprehensive resource for researchers and policymakers alike.

References

Bia, M., Blanco, G., & Valentova, M. (2025). The heterogeneous impact of parental leave take up on the wage distribution: Evidence from Luxembourg. *Econometric Reviews, 44*(3), 356–383. https://doi.org/10.1080/07474938.2024.2417167

Bia, M., Flore-Lagunes, A., & Mercatanti A. (2022a). Evaluation of language training programs in Luxembourg using principal stratification. *Observational Studies, 8*, 1–44. https://doi.org/10.1353/obs.0.0000

Bia, M., Mattei, A., & Mercatanti, A. (2022b). Assessing causal effects in a longitudinal observational study with 'truncated' outcomes due to unemployment and nonignorable missing data. *Journal of Business and Economic Statistics, 40*, 718–729. https://doi.org/10.1080/07350015.2020.1862672

Biewen, M., Fitzenberger, B., Osikominu, A., & Paul, M. (2014). The effectiveness of public-sponsored training revisited: The importance of Paul data and methodological choices. *Journal of Labor Economics, 32*, 837–897.

Bousselin, A., & Solaz, A. (2026). Children's perceptions of economic hardship: Measurement, mechanisms and well-being. Aubervilliers: INED, (Documents de travail INED; no. 307).

Data Dictionary: https://igss.gouvernement.lu/dam-assets/microdata-platform/data-dictionary.pdf.

Docquier, F., Gordan, A., Tenikue, M., & Uljarevic, A. (2024). Attitudes towards immigration in a highly multicultural society: The roles of foreign background and local exposure. Esch-sur-Alzette: LISER, 54 p. (Working papers; no. 2024-08).

Lechner, M., & Wunsch, C. (2013). Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Economics, 21*, 111–121.

LMPD: Luxembourg Microdata Platform on Labour and Social Protection—Inspection générale de la sécurité sociale//Le gouvernement luxembourgeois. https://igss.gouvernement.lu/fr/microdata-platform.html

Moran, V., Bia, M., Thill, P. et al. (2024). The impact of patient registration on utilisation and quality of care: a propensity score matching and staggered difference-in-differences analysis of a cohort of 16,775 people with type 2 diabetes. *BMC Primary Care, 25*(1), 254. https://doi.org/10.1186/s12875-024-02505-2

Thomas, F., & Zanardelli, M. (2023). Methodologcal note. *Social Security*. https://igss.gouvernement.lu/fr/publications/apercus-et-cahiers/cahiers-methodologiques/202305.html

Valentova, M. (2019). The impact of parental leave policy on the intensity of labour-market participation of mothers. Do the number of children and pre-birth work engagement matter? *Journal of European Social Policy, 29*(3), 428–445. https://doi.org/10.1177/0958928718776826

Valentova, M. (2024). How do parents care together? Dyadic parental leave take-up strategies, wages and workplace characteristics. *Work, Employment and Society*. Advance online publication. https://doi.org/10.1177/09500170241229281

Valentova, M., Amjahad, A., & Genevois, A.-S. (2022). Parental leave take-up and its intensity. Do partners' workplace characteristics matter? *Journal of Social Policy*. https://doi.org/10.1017/S0047279422000885

Michela Bia is a Researcher at the Luxembourg Institute of Socio-Economic Research (LISER) and Affiliate Professor at the University of Luxembourg, teaching a course on Cross-sectional Data and Big Data. She holds a Ph.D. in Applied Statistics from the University of Florence and has been a Visiting Researcher at UC Berkeley and the University of Miami. Her research focuses on programme evaluation, causal inference, machine learning and labour markets. Her work has been published in *The Journal of Business and Economic Statistics*, *The Journal of the Royal Statistical Society: Series A, Econometric Reviews, The Stata Journal* among others.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 20
Access to Administrative Data in Malta



Rose Marie Azzopardi

Abstract The situation in Malta regarding access to data is currently in flux; there is no standardised system through which a researcher might gain access at present, but legislative processes and new institutions, such as the Public Administration Data Strategy and the Data Governance Council, are expected to facilitate access within the public sector in the future. Perspectives gleaned from the producers of administrative data suggest that there was no standardised way through which to request data, except for some entities whose websites provided a formal request form. Survey replies from the public administration indicated that several departments (where possible) did provide information for research purposes; however, there was a feeling of unease due to issues of privacy, traceability and unknown motives for such requests for data. Thus, the situation proceeds on a case-by-case request and approval at present. Concerning academic uses, most academics have neither made use of administrative data nor encouraged their students to do so. Many of those who did use such data found it generally difficult to gain access thereto. The Data Strategy, and accompanying Malta data portal, could change such access, but information regarding what type of microdata can be made available may still limit such access due to anonymity issues associated with a small population. Only time will tell if access to administrative data has improved in reality.

1 Introduction—Setting the Scene

> Within Government, data is central to the delivery of public services. We need to ensure that this data is open for its appropriate consumption and use, empowering society to become data-driven. (Sultana, 2023, p. 4)

The aforementioned quote comes from the Principal Permanent Secretary in his forward to the Public Administration Data Strategy document for 2023–2027. According to him, the Strategy 'views data as a national asset…bears inherent

R. M. Azzopardi (✉)
University of Malta, Valletta, Malta
e-mail: rose.m.azzopardi@um.edu.mt

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_20

risks…needs to be governed, protected, shared and consumed within a comprehensive regulatory framework…unlocking the value of public sector information' (ibid., p.5).

This chapter focuses on Malta and how the issue of data and access thereto is dealt with by researchers, academics and students. While there is a wealth of data that is gathered daily, and the law puts the onus on data collectors to keep registers of the information, there is no systematic presentation of such data or even a structured and standardised way in which such data is kept, managed or even offered for research purposes.

Administrative data contains information that is not primarily intended for statistical purposes; however, the new focus on private and public data, which can be utilised for policy and strategy, has refocused attention on huge amounts of data that is gathered by both government and private entities. From the public sector, these can include taxation, social security, health, education, labour market, licencing and permits systems and different types of registers. From the private sector, information regarding clients, sales, customer cards, utility companies' details, telephone directories, etc. have been collected and analysed to inform the seller about the best way forward. However, such data is not intended for publication or for official statistical use; rather, it is a wealth of information that may not be fully utilised for research purposes which might then, in turn, inform policy. Some of this data is, however, being utilised by the national statistics office for official publication; this notwithstanding, such data remains under the ownership of the entity (whether public or private) that initially generated such data.

Apart from the introduction, this chapter is divided into five sections: the first looks at both legislative and institutional changes in Malta that follow from EU legislation; the second presents the methodology used to collect information about access to administrative data in Malta; the third provides responses from the government side to a dedicated survey that we ran for the present book's purposes and examines expected changes in the near future; the fourth focuses on the experience of academics with access to administrative data and the last section concludes.

2 Legislative and Institutional Framework for Data

There are several institutions whose role is linked to either data collection or protection. These include the Malta Digital Innovation Authority, the Malta Information Technology Agency, the National Statistics Office and the Information and Data Protection Commissioner.

2.1 The Malta Digital Innovation Authority (MDIA) was established in 2018 and acts as a regulator in the sector and encourages innovative technologies. Its website states:

> The European Data Governance Act (DGA) is a game-changer for the EU's data-driven economy. Malta, a champion of digital innovation, is taking a leading role in implementing this framework.

The MDIA plays several roles in the process within areas such as data intermediation services and data altruism organisations. Notifications of the former and registration of the latter need to be made to the MDIA as the relevant local authority. Supervision, as to adherence to the DGA requirements, is also carried out by MDIA.

The MDIA is also involved in the European Data Innovation Board, which was established by the European Commission. The Board's objective is to facilitate the sharing of best practices in the areas under the DGA, including data intermediation, data altruism, the use of public data which cannot be provided as open data and interoperability standards across different sectors.

2.2 The Malta Information and Technology Authority (MITA) 'operates a Tier III Data Centre which hosts the vast majority of government information systems, including those critical for the operation of the Maltese Public Administration' (MITA, 2024). Its website also lists both the new Malta Data Portal (the national Open Data Portal) and the Internal Data Sharing Services. The Open Data Portal is a 'one-stop shop for discovery, viewing, downloading and on-line usage of data,' which will be made available publicly to all, while the latter 'will be restricted to Public Administration Officers and other authorised persons.' This is in line with the Public Administration Data Strategy 2023–2027, whereas data management is conducted through the Corporate Data Repository (CDR) (OPM, 2023). This centralised approach to accessing common data from various governmental departments has several benefits: ease of use, sharing of resources and avoidance of duplication.

The Malta Public Administration Data Strategy 2023–2027 aims to provide for better and more efficient use of the significant amount of data generated by the public sector. The Action Plan should unfold in the coming months. The strategy document states that:

> Unfortunately, the large availability of data has also resulted in a fragmented data ecosystem within the Public Administration, where data management is not adopted in a holistic manner. (OPM, 2023, p. 9)

2.3 Another local institution is the National Statistics Office, which was established through the Malta Statistics Authority Act XXIV of 2000, replacing the Central Office of Statistics (Laws of Malta, 2000). Its role is to 'collect, compile, extract and release official statistics.' Subsidiary legislation 422.01 (Legal Notice 28 of 2014) on 'Fees Payable to the National Statistics Office for the Processing of Data Requests Regulations,' indicate the fees payable on an hourly basis for the NSO to process customised requests for general users and students, and in the form of data, advisory services or professional services (Laws of Malta, 2014). Another relevant subsidiary legislation 422.02 (Legal Notice 368 of 2020) 'Requesting and Sharing of Data Regulations,' stems from the MSA Act (Laws of Malta, 2020). It gives the right to the Director General of the NSO, who 'may require any person, undertaking or public authority to provide him with information and registries the Director General

deems necessary in order to contact persons and undertakings for the purpose of obtaining information for statistical purposes.' Furthermore:

> …at his sole discretion, for official statistical purposes only and when no other possible alternatives are available, share non-financial data, as needed, from the Business Register, with public entities, and he may impose such conditions as he may deem necessary for the sharing of the information. (NSO, 2024)

A third subsidiary piece of legislation No 422.03, Legal Notice 261 of 2022, provides the NSO with the possibility of 'using its resources (to) carry out statistical work for other public entities against a cost recovery charge' (Laws of Malta, 2022). Safeguards are in place that ensure the protection of individuals so that they cannot be identified.

2.4 The Information and Data Commissioner is the national supervisory authority which monitors and enforces provisions within three pieces of legislation. The Information and Data Protection Commissioner's role is to ensure that the GDPR, transposed nationally into the Data Protection Act, is adhered to. The Office is responsible also for the Freedom of Information Act (Laws of Malta, 2009). While the latter legislation ensures transparency and access to data, the individuals whose data is being provided are protected through anonymity and non-traceability. Legislation that encourages greater data sharing services is expected to put structures into place that provide access to data that was also collected for research purposes.

3 Methodology—Gathering Our Data for This Chapter

The methodology to gather information on the current situation in Malta was based on various sources to try to find the different pieces to the data access puzzle and to elicit different viewpoints on the topic. The study started out with desk research on the institutions dealing with data. This was presented in the previous section. A questionnaire was then sent to all academics at the University of Malta (UoM) and the Malta College for the Arts, Science and Technology (MCAST), to enquire about their experience with administrative data in their research. Those targeted included current and potential users desiring easier access to data. In the case of UoM, a mailshot was sent to all full-time and part-time academics, with the assistance of the Communications Department. Unfortunately, the response rate was not very encouraging. A second email was sent to the Deans of the fourteen faculties to request their assistance in drawing attention to the questionnaire. This elicited a higher response. In the case of MCAST, an email was sent to the principal, who forwarded the request to the relevant department and the link to the questionnaire was sent to the academic staff. Unfortunately, in this case there was only one response, which was excluded from the responses and, thus, the analysis focuses solely on the responses received from the University of Malta.

The communications coordinators of all Government Ministries in Malta were then contacted with a set of questions regarding their Ministry's role as providers

of administrative data. Some of the coordinators requested clarifications; others sent the request on to relevant officials. An online meeting was held with the Chief Data Officer, who explained the Data Governance Council's framework. Another meeting was held with the regulator, the Information and Data Commissioner thereafter.

The third set of questions was sent to the National Statistics Office and sought to understand the role of administrative data for official published statistics. The NSO is possibly the primary user of such data and the law provides it with the right to request administrative data, as discussed in the previous section. The main result of these questions was to confirm the existing legal and institutional framework as described above.

4 Access to Administrative Data in Malta—The Producers' Perspective

4.1 The first interview conducted was with the Government's Chief Data Officer (CDO). The discussion rested on the two pieces of relevant legislation. The DGA put into place requirements to appoint Competent Authorities and, in line with the Act, the Malta Digital Innovation Authority was designated as the competent entity for data altruism and data intermediation. Existing enacted laws are being reviewed so that a holistic Framework may be sustained and to comply with these acts and directives. For instance, the role of Chief Data Officer was created recently so as to facilitate Data-related matters; similarly, the Data Governance Council (now Chaired by the CDO) was created through the transposition of the EU Open Data Directive, and resulted in Chapter 546 of the Laws of Malta and its related subsidiary legislation (Laws of Malta, 2024). Local legislation is now being reviewed to ensure legal provisions are not in conflict.

Similar to the EU's 'data.europa.eu' website, Malta is also in the process of populating the Maltese version, with the relevant information on registers and databases, with the aim of launching the finalised enhanced version by the end of 2024 (Government of Malta, 2024). At the time of writing, the website's structure and framework was available (albeit in Beta format). In effect, the website is aimed to be a type of resource which offers a 'register of registers.' As per the EU's Data Act, it focuses on open data and protected (closed data) metadata, rather than microdata (European Commission, 2024). The portal publishes solely metadata (such as owner, status, classification) in a machine-readable format about the registers. Datasets marked as open by the owners of such data, will provide a link to get such data from the source. For closed data, a request to the data owner would still need to be made. The Chief Data Officer's team remains committed to assist data owners to open up their data where possible and also assist with any data related queries.

The ongoing process includes a compilation of the list of registers that derive from the Maltese legislation, including the Malta Business Register, the Person Register and several administrative registers. The website's disclaimer states:

> This version of the Malta Data Portal is designed to enable a shared platform for the management and support of metadata about Registers and Datasets emanating mainly from the laws of Malta and EU Regulations. The contents of this portal are currently works in progress and should not be considered as providing official records unless such records are confirmed through a confirmation request. (ibid.)

A non-exhaustive list of such registers (but not the contents) can be found on the website, which proclaims 404 registers, 243 datasets and 52 publishers. These derive from legislative acts and subsidiary legislation.

The metadata will be provided for thirteen sectors: Education, Culture and Sport; Environment; Population and Society; Government and Public Sector; Agriculture, Fisheries, Forestry and Food; Economy and Finance; Regions and Cities; Energy; Justice, Legal System and Public Safety; Health; Transport; Science and Technology; and International Issues. The highest number of datasets are those for the Environment followed by Public Administration. The website continues to be populated (The above describes the situation as on 16 December 2025).

According to the Chief Data Officer, there is no consistency as to the type of data, its quality or comprehensiveness about Administrative Registers at present: an exercise is currently underway to standardise these so that they may be published on the enhanced data portal. As regards both the DGA and the Data Act, the EU itself is still developing the tools to be able to assist Member States in providing certain categories of sensitive data (concerning health, for example) in an anonymised manner. Furthermore, there is significant work that needs to be accomplished first, since administrative data is updated daily and, thus, can change from one day to another. The website also has another section with relevant resources.

4.2 The discussion with the Commissioner for Information and Data Protection centred mainly on the GDPR, the Freedom of Information Act and the process followed for a request to be handled; most of these requests are initiated by journalists. Micro-level data is mainly requested for official use. However, since Malta does not have a population register, other sources (such as passport or identity cards) are used as person identifiers. The quality of the registers is not always ideal. There are also restrictions in place and one needs to make a distinction between directly identifiable data and indirectly identifiable ones. In such cases, certain details remain suppressed. The Commissioner stated that there were guidelines regarding anonymisation techniques.

4.3 The responses received from the Ministries varied. All public entities generate some type of administrative data. Not all entities said that their data is requested for research purposes. In most cases, while it is easy for the person making the request, some departments complained that this put extra work on desks of the entity's staff, not only because data needed to be anonymised, but also because the data requested might not always be organised and might need to be collated from different sources. Certain bodies said that when the request does not come through a parliamentary question or the Freedom for Information Act, then the requestor needs to state why the information is needed.

Certain entities, such as Jobsplus (the public employment agency) has a research policy in place that needs to be followed by anyone requesting data for research purposes. Researchers can also request face-to-face interviews (for example, with those registering as unemployed). After consent is given by the individual, the Labour Market Analysis Unit provides the details to the interviewer. According to the Employment and Training Services Act Cap 594, Jobsplus is obliged to gather data and to either use it itself or even transfer it to third parties, in compliance with the GDPR and the Data Protection Act. Personal data is never disclosed to third parties.

The entities were also asked if the requestors provide feedback emanating from their research for policy purposes. Some entities do ask for a copy of the research, which can then be utilised internally, and some also request a disclaimer. For example, any research/dissertation using Jobsplus data needs to state that:

> The responsibility for the opinions and conclusions expressed in this publication rests solely on the author(s), and the publication does not constitute an endorsement by Jobsplus of the questions, opinions and conclusions expressed in it.

There were a few governmental bodies who stated that nobody had ever asked for their data for research purposes, and that their data was only used internally for 'record-keeping, forecasting, planning, etc.' Some indicated that data is often requested officially by EU institutions. These entities were fearful of 'the extent and potential risks of having such data divulged.' One entity stated that all of the data that it is obliged by law to publish finds its way into the entity's annual report, the findings from which are made public after they are presented to Parliament.

There were a few entities who maintained that their data was sensitive and could also be commercially sensitive and, therefore, they remained wary of granting access to their data, particularly due to fear of ulterior motives behind such requests.

Many entities were willing to aid researchers in their work, particularly if it could 'be shared without compromising privacy or security.' Some entities ensure that this is reviewed by the giver of the data and information before any research is published, so as to guarantee no breaches occurred and that no person could be identifiable. Overall, many entities were positive about the process of providing anonymised data for research purposes, although this entailed extra work on the part of the provider.

5 Academic Experience with Access to Administrative Data—The Users' Perspective

5.1 The number of academics who replied to the questionnaire reached 402. They hailed from different faculties, institutes and centres and the years at the academic institution ranged from 3 to 50 years. Females accounted for 41.8% of respondents.

The majority of those who replied (63.6%) indicated that they had never made use of administrative data. The reasons for this varied, such as: lack of data, time-barred for historical research, they were outrightly refused, too difficult so finally gave up,

access was complicated, no data available and unaware of administrative data or how to access it. One answer stated that it was due to 'GDPR issues and Data Protection clearance which have gone beyond and above any ethical considerations and have taken an extraterrestrial legal life of their own' (Respondent 233).

When asked if they would be interested in utilising such data, should this become available, the majority 54.8% said yes, while another 11.9% said maybe. Only 7.1% replied negatively, while a good 26.2% replied that they were not sure what kind of administrative data would be available in their field of research.

A significant 62.8% of 'No' respondents said that they never encouraged their students to use administrative data for their dissertations and research. A small minority (2.3%) actively discouraged them from trying to make use of administrative data. Almost a third (32.6%) of 'No' respondents said they had encouraged students to use such data, with a further 2.3% themselves helping to contact government entities and even private companies.

Most academics agreed that it was difficult or very difficult to access data. There were 36.4% of respondents who had replied in the affirmative, of having made use of administrative data. The areas ranged from data on the labour market, medical field, health, education, energy and renewable energy, education, agriculture, social well-being, research and innovation, gerontology, taxation, business and spatial planning.

Around a third (34.4%) of 'Yes' respondents said that it was easy or relatively easy to access administrative data. Many (37.9%) found it either difficult or very difficult to access administrative data. The remainder (27.7%) said that they had access, but had also encountered some difficulty.

One question asked academics how they gained access to administrative data. This allowed them to choose between contacting the Ministry's communications officer, the research unit of the entity, making a formal request online, through a parliamentary question or any other means not mentioned in the questionnaire. In terms of ways of requesting data, 35.7% said that they had directly contacted the entity's research unit, while another 25% made a formal request online. A further 14.3% contacted the communications unit of the Ministry or entity. Some respondents used more than one method to access data. Some went through websites or official annual reports to devise their own dataset. A small percentage requested data, which was not initially made available, via a parliamentary question.

Data was made available in all cases under condition of anonymised data. In certain cases, it could only be used for that particular research project and must have already gone through the ethics approval process of that academic institution. When asked if metadata was available in their area of research, the range of answers were varied: no, yes, sometimes, not always, not in a structured manner, available but not made available to the researcher, not easily made available and not available in all areas of analysis.

When asked if they had the possibility of linking different sets of data on an individual level, only 17.2% of 'Yes' respondents replied in the affirmative. A significant portion, 65.5%, said that this was not possible. A few replied that while the possibility existed, getting permission was a 'nightmare.' A few more replied that this is sometimes difficult as it means additional work for the people who needed to link

the datasets together to present the researcher with one anonymised database. Some researchers were unaware of this possibility of linking different datasets.

Not all the 'Yes' respondents had encouraged their students to utilise administrative data. In fact, 61.1% had encouraged them, but the remainder had not done so. The reasons were that this was sometimes a lengthy process and students did not have the time to do so, or the process was too cumbersome and did not always have a positive outcome; to that end, time would have been wasted waiting for data that, in the end, was not made available.

The researchers were asked for the advantages that they saw in using administrative data. These were various and included: detailed information, useful, needed to see trends, useful for policy-making, reflecting the real situation and changes over time, efficient use of ready-made data, reduces the need for expensive surveys, provides the possibility to link many variables, secondary data is economical, readily available and does not need to be collected, complete, gives a wider picture, it represents the population not a sample, evidence at the local level, presents a reasonably up-to-date overview, quick access to data and such data provides researchers with the ability to plan and base projects and research on real data.

Some of the lengthier replies included: 'It generates value from otherwise hugely underutilised data,' and 'the availability of large volumes of data in a short amount of time and the possibility of rapid and effective anonymisation' (Respondents 231 and 59 respectively).

The academics were then asked if they saw any problems with using administrative data. A vast majority did not see any problems with its usage. Others mentioned comparability and reliability over the years. Some mentioned the risk of data gaps; others mentioned the lack of access to specific data, a few mentioned the lack of knowledge about how accurately the data is collected, while others lamented that there were verifiability problems. A few referred to such data being misused or misinterpreted.

When asked if access to administrative data were to be made easier how this would impact future research projects, 88.9% replied positively, while 2.8% said they were not sure, and the remainder 8.3% said they did not believe that it would make future research projects more possible.

The academics were then asked to suggest best practices for the research community to have open data access, while still fulfilling the privacy rule. The suggestions included: anonymised data to be made available on a shared platform; one person in each entity whose role is for research purposes; a dedicated unit in large entities dealing solely with data and its access; reduce the 'bureaucratic labyrinth' for researchers; an index of available data for research; MOUs between institutions to allow access; data to be anonymised by producer to eliminate risk of a breach in GDPR; ethics training for staff collecting/producing data; a centralised data request centre; non-identifiability is key; clear trail of responsibility in case of a breach; more information for academics regarding data access; ensure confidentiality; create databases to be used for research; password protected access which is also traceable and the use of the University's drUM (data repository of the University of Malta). There were also remarks regarding data that is not sensitive and that could be

made 'available in the public domain and regularly updated.' There were mentions of certain websites that used to have data (on education and school statistics and foreign workers in the labour market, for example) and which are no longer available. Some academics mentioned best practices from abroad, such as official data centres where requests can be made accompanied by a study protocol and shortening the process where ethics approval is not needed in cases where there is no risk to individuals.

The final question asked for any other comments that respondents wanted to make. These included: the fact that it was a shame that there was such a wealth of data which was not being capitalised on for research purposes; that lengthy ethics approval even when anonymised data is available is disheartening and delays research output; data financed by public funds should be made available free of charge; data gathered by private companies which have a national impact, such as weather and climate data, should be free or subsidised by government for research; more awareness about data repositories and that entities do not want to waste time producing anonymised data for researchers. One academic even suggested that the more open the data, the more accountable and transparent governments need to be in their public actions.

5.2 From my own personal experience with access to administrative data for research purposes, the situation is different across the diverse governmental entities. Some of them do have online request forms, such as labour market data from the governmental employment agency. With such institutions, accessing data is only limited by the actual data available, the possibility of providing anonymisation within such a small country and (in some cases) financial concerns need to be taken into consideration when such data is provided for an administrative fee. Perhaps the most problematic is the second one, in terms of anonymity. For example, on one EU project, in order not to be able to identify persons, Eurostat in the case of Malta could only provide microdata at the two-digit level under the ISCO classification, as anonymity would not have been guaranteed at the four-digit level. This was a decision taken in the case of Malta only and this is related to the small size of the population where there may only be one person who would qualify as a specialised surgeon in a particular area or as a professor within a particular domain. If such data were to be made available, then the specific person would be easily identifiable. Certain governmental entities are more approachable than others. I have found collaboration on areas in which I conduct research, such as the labour market, green jobs, and education. In certain cases, a request for the NSO to gather information for research through questionnaires was also possible; for example, in the case of single mothers where anonymity was guaranteed, even if this was against payment.

5.3 The final remarks regard the role of the University and its academics in promulgating data access through their own research and agreements reached with other entities. First, the University of Malta launched its data repository 'drUM,' the purpose of which is to 'collect, archive, organize and when possible, disseminate data produced by UM researchers.' At the time of writing, the website had 18 items and included datasets utilised for papers and research conducted by its academics (University of Malta, 2024). Second is the University's Open Access Repository that showcases academics' published research, which include several papers based or related

to administrative data, such as Zuccarelli et al. (2021), Eldrup and Debus (2020), Azzopardi Muscat et al. (2017) and Borg (2019); these works mostly concern the health and labour market sectors. Third, there is also collaboration by academics and law enforcement bodies which has set up a dedicated website, CrimeMalta (Formosa, 2024). Lastly, the University of Malta has set up several platforms for research, based on different topics, which also collaborate with civil society, NGOs and government entities to base their research on existing and unpublished information and data, for example, concerning the number of third country migrants seeking legal aid services provided by NGOs or the number of women who seek assistance because of domestic abuse.

6 Conclusion—Differences in Outlooks

Malta has a new strategy to ensure that data can be provided to the general public. This should help researchers in their work. However, works are still in progress both in terms of the data portal as well as on the standardisation and comparability processes across the different databases. The civil service appears willing to push the agenda forward. Positive developments include the data portal for government in Malta, which is currently under construction, and the launch of the person register on 16 May 2024, which can be accessed by all government entities, thereby providing the necessary links between the datasets.

Several government entities replied to the questions sent regarding the types of data that they collect and most entities maintained that, when asked, they provided data for research purposes. While the responses from the governmental bodies varied, concerning the process by which data is requested, the type of data that can be made available, the amount of requests received and the work that this entails: nevertheless, the majority of entities were willing to help researchers in their work. The main concerns remain privacy, security, identifiable details and the capacity to anonymise data.

However, most academics who filled in the questionnaire stated that they generally found it difficult to access administrative data. The data portal might aid in this regard, even though what type of access to microdata will become available is still unknown. Furthermore, this research has shown that it is still early days concerning data availability for research purposes and several milestones need to be reached in this process in order for it to be sustainable. A particular concern is the issue of anonymity, which can be problematic from a small-scale perspective. Certain microdata could lead to the traceability of the person involved, due to the size of the population; for example, the lack of mass of expertise in the labour market and, thus, the limited persons in each specialisation; the type of rare disease suffered by someone and the type of services offered by a business or the type of product being manufactured. Traceability remains the main problem within a small population.

References

Azzopardi, M. N., Calleja, N., Buttigieg, S., & Sherry, M. (2017). Malta: Health system review. *Health Systems in Transition, 19*(1), 1–37.

Eldrup, N., & Debus E. S. (2020). *International validation of the population based Malta vascular registry: A vascunet report*. https://www.sciencedirect.com/science/article/pii/S2666688X20300253

European Commission. (2024). https://commission.europa.eu/news/data-act-enters-force-what-it-means-you-2024-01-11_en.

Formosa S. (2024). CrimeMalta. https://www.um.edu.mt/projects/cloudisle/DATA1/crimemalta/about.html

Government of Malta. (2024). *Malta data portal*. Available at version https://open.data.gov.mt/dashboard.html

Information and Data Protection Commissioner. (2024). https://idpc.org.mt/

Laws of Malta. (2000). *Malta statistics authority act*. Chapter 422. https://legislation.mt/eli/cap/422/eng

Laws of Malta. (2009). *Freedom of information act*. Chapter 496. https://legislation.mt/eli/cap/496/eng

Laws of Malta. (2014). *Fees payable to the national statistics office for the processing of data requests*. https://legislation.mt/eli/sl/422.1/eng

Laws of Malta. (2020). *Requesting and sharing of data regulations*. https://legislation.mt/eli/sl/422.2/eng

Laws of Malta. (2022). *Tax exemption and statistical work for public entities regulations*. https://legislation.mt/eli/sl/422.3/eng

Laws of Malta. (2024). https://legislation.mt/eli/cap/546/eng and https://legislation.mt/eli/sl/546.1/eng

MITA. (2024). Data Center and ICT Services – MITA (gov.mt). https://mita.gov.mt/portfolio/data-center-ict

National Statistics Office. (2024). *National and European legislation*. https://nso.gov.mt/legislation

Office of the Prime Minister (OPM). (2023). *Public administrative data strategy 2023–2027*. https://mita.gov.mt/wp-content/uploads/2023/07/Public-Administration-Data-Strategy-2023.pdf

Sultana, T. (2023). *Forward to the public administrative data strategy 2023–2027*. https://mita.gov.mt/wp-content/uploads/2023/07/Public-Administration-Data-Strategy-2023.pdf

University of Malta. (2024). *Data repository of the University of Malta*. https://drum.um.edu.mt/

Zuccarelli, M., Micallef, B., Butler, D., Serracino-Inglott, A., & Borg, J. J. (2021). *Improving the data quality of spontaneous ADR reports: A practical example from Malta*. https://www.tandfonline.com/doi/abs/10.1080/14740338.2022.1993820

Rose Marie Azzopardi, D.P.A., B.A., M.A., Ph.D. is Associate Professor of Economics at the University of Malta, with a research focus on the labour market. Her first two degrees are from the University of Malta, while her Masters and Ph.D. are from the University of Sussex, UK. She has also followed short courses at Victoria University, Australia, and Harvard Business School, USA. Her other research interests include small states, migration and regional integration. She has presented papers at global conferences and published in several international journals. She is the national expert on several EU programmes and projects. She has conducted research and engaged in consultancy work with the Government of Malta, the United Nations, World Bank, European Union institutions and the Commonwealth Secretariat. Her current research on the Future of Skills is sponsored by the HSBC Foundation.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 21
Statistics Netherlands Microdata: A User Perspective



Ivo Gorissen, Maarten Bloem, Vincent van Hees, and Ran van den Boom

Abstract Statistics Netherlands (CBS) enables authorised users to conduct a wide variety of research via its secure Microdata environment; it provides users with the choice of all CBS datasets that are, potentially, enriched with external datasets. The challenge this presents is to keep meeting the international research community's ever-changing requirements when using the Microdata facility.

1 Introduction

CBS Microdata Services provides a wealth of detailed information for researchers. In 2024, 1900 researchers from 225 institutions conducted their research with CBS Microdata. Close to 7500 datasets are available for researchers; there are also an additional 1500 datasets that have been brought in by researchers themselves in order to enrich the CBS Microdata.

Microdata are individual-level data collected by CBS that provide insights into various aspects of social, economic and demographic phenomena. This chapter explores the possibilities and limitations of working with Dutch Microdata through the CBS Microdata Facility. First, we discuss the conditions for access, how research is made possible and describe the costs involved. We then turn to both understanding and exploring the Microdata's structure. This is followed by an examination of the data available to researchers and a discussion of a selection of research papers based

I. Gorissen (✉) · M. Bloem · V. van Hees · R. van den Boom
CBS, The Hague, Netherlands
e-mail: i.gorissen@cbs.nl

M. Bloem
e-mail: m.bloem@cbs.nl

V. van Hees
e-mail: v.van.hees@cbs.nl

R. van den Boom
e-mail: r.vandenboom@cbs.nl

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_21

on CBS Microdata research. Finally, we outline some pitfalls and challenges and shed light on future developments.

2 Accessing CBS Microdata and Costs

2.1 Authorisation for an Institution

CBS Microdata is accessible for researchers only. Research institutions, or so-called 'statistical enclaves,' need to apply for access to CBS Microdata. If they meet the conditions,[1] then they will be authorised for a limited period of time (three years maximum), after which time an extension is possible. They agree to data usage terms and undergo a vetting process in order to ensure compliance with data privacy and confidentiality regulations.

2.2 Working on a Research Project

Users can explore available datasets in the online catalogue and can select those relevant to their research projects or analyses. A full overview of the datasets provided is available on the CBS Microdata website.[2] Although parts of the website are available in English, the metadata is available almost exclusively in Dutch; however, expert meetings held with the researchers at the beginning of a project can be done in English. CBS experts provide assistance to the researchers regarding their choice of datasets during the expert meeting. A separate project is created on the CBS Microdata environment for each research question. The accompanying project request is assessed on feasibility and GDPR compliance. At this stage, CBS also checks that the data will not be misused for either control or enforcement purposes. The CBS Ethical Committee is consulted, where necessary.

Researchers can access the Microdata via a connection on their own computer to the CBS secure Microdata environment once the authorisation and safety procedures are completed. The data is safely stored and kept on CBS servers, but can be conveniently prepared and analysed from the researcher's computer.

[1] The main criteria for admission are that the institution either has the conduct of statistical or scientific research as its principal activity or has an enclave entirely devoted to scientific research and publishes the results of this research for the public. In addition, the institution must have legal personality and be located within the European Economic Area (Member States of the European Union plus Liechtenstein, Norway and Iceland) or in a country with an adequacy decision. More information can be found in the Policy rule on access to CBS microdata (https://wetten.overheid.nl/BWBR0045454/2021-08-01; Dutch only). For more information about the adequacy decision see: Data protection adequacy for non-EU countries.

[2] www.cbs.nl/microdata.

Researchers are given the possibility of importing additional datasets, if they need them for their research and provided that they are legally entitled to use these data and the applicable privacy and data protection legislation is respected. CBS will pseudonymise the datasets in such a way that they can be linked to CBS Microdata.

Finally, researchers can add datasets from other institutions that have made some of their datasets available for linkage (e.g., Lifelines, LISS Panel, PIAAC and SHARE[3]). These data can also be linked to CBS Microdata.

2.3 Costs

Since providing access to the Microdata is not being funded by the Government, CBS charges the users for the services that involve the provision of access to the data. Depending on the intensity of use, the costs for a project can vary from 3500 euro to 20,000 euro or more (for a long-lasting, larger research project).

2.4 Data Security and Confidentiality

Ensuring data security and confidentiality is paramount, given the sensitive nature of individual-level data. Users of CBS Microdata must adhere to strict protocols in order to prevent unauthorised access, maintain data integrity and safeguard respondent privacy. This includes implementing encryption measures and complying with relevant data protection regulations.

A so-called output check is conducted in order to export (draft) results from the secure environment. All results that leave the secure CBS Microdata environment are subjected to this; in doing so, CBS ensures that results leaving the CBS Microdata environment are not traceable to individual persons, businesses, or other (semi) public institutions, such as schools or hospitals.

[3] Lifelines is a large, multigenerational cohort study with health-related data from the northern population of the Netherlands; the LISS panel (Longitudinal Internet Studies for the Social Sciences) is an online household panel; PIAAC (Program for the International Assessment of Adult Competencies) is an OECD survey of adult skills; and SHARE (Survey of Health, Ageing and Retirement in Europe) is a multidisciplinary and cross-national panel database of microdata on health, socio-economic status and social and family networks.

3 Understanding and Exploring the Microdata Structure

CBS Microdata is structured at the individual level, historically providing granular information about survey respondents or administrative registers. The main source of Microdata has shifted from surveys to administrative registers over the years. More recently, scraped internet data has become a third source of Microdata.

Understanding the structure of the Microdata is crucial for formulating research questions, designing analyses and interpreting results accurately.

Before conducting analyses, users often need to preprocess the Microdata in order to ensure quality and consistency. This may involve tasks such as:

- Data Cleaning: Identifying and handling missing values, outliers and inconsistencies within the dataset;
- Variable Recoding: Converting categorical variables into numerical formats or aggregating categories for easier analysis;
- Normalisation: Standardising variables to a common scale in order to facilitate comparison and interpretation;
- Subset Selection: Choosing relevant subsets from among the data based on research objectives or on specific criteria.

Exploratory Data Analysis (EDA) is a crucial, subsequent step in understanding the characteristics and patterns within CBS Microdata. Techniques such as summary statistics, data visualisation and correlation analysis can provide insights into relationships between variables, distributional properties and potential trends. EDA helps users to formulate hypotheses and to refine their research questions before proceeding with more advanced analyses.

CBS Microdata enables a wide range of statistical analyses that can address either research questions or policy concerns. Users may employ various methods using statistical software like R, SPSS, STATA etc., depending on the nature of the data and research objectives, including:

- Descriptive Statistics: Summarising key features of the data through measures such as mean, median, standard deviation and frequency distributions;
- Inferential Statistics: Drawing conclusions and making predictions about populations based on sample data using techniques such as hypothesis testing and confidence intervals;
- Regression Analysis: Examining relationships between dependent and independent variables in order to identify factors influencing outcomes;
- Clustering and Classification: Grouping individuals based on similarities or classifying them into predefined categories using techniques like cluster analysis or decision trees.

4 Overview of the Data

The data available in CBS Microdata[4] can roughly be distinguished in data on citizens, on firms and concerning environmental data. Each of these data sets has a main encrypted key: personal identifiers, firm identifiers and addresses. Research between these fields is possible using datasets entailing two of these encrypted keys (e.g., where there is a dataset containing information on individuals and their addresses (both encrypted) and there is a data set on individuals working in a firm).

4.1 Data on Citizens

Most datasets on citizens are available at CBS Microdata. These data vary from population data to income, health and crime.

Population

Researchers can find administrative data about the Dutch population from October 1994 onwards under this section, including month and year of birth, sex, migration background and household information (e.g., a person's position in the household, marital status, adoption, etc.).

Several datasets based on the so-called Population Network are available in the section on Population. This data presents a whole population network file that has been constructed from administrative data. The network incorporates over 1.5 billion relationships between all 18 million inhabitants of the Netherlands. Relationships are identified between individuals who live in the same household, live close to each other, work for the same company, attend the same educational institution, or who belong to the same extended family.

Labour and social security

Administrative data on jobs and wages of individuals working in the Netherlands and a wide variety of information about social security benefits, welfare benefits and pensions are made available within this realm. These data can, for example, be combined with the Labour Force Survey (EBB) or the National Working Conditions Survey (NEA).

Income and expenditure

Administrative data on the income of persons and households are paralleled by information about expenditures from the Budget survey. CBS has also developed an administrative dataset showing the socio-economic category of each citizen in each month, based on their income and wealth and a dataset showing the socio-economic

[4] Catalogus microdata | CBS (Dutch only). For an overview of the subjects see: https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research/microdata-catalogue.

status scores of households based on their financial wealth, educational attainment and recent employment history.

Education

This section includes administrative datasets about all levels of education of pupils' and students' educational participation and outcomes (such as diplomas and exam results) and data on schools and education personnel. A collection of datasets is also available that allows files to be easily linked together, thereby making it possible to track cohorts of primary and secondary school students over time using data on students' education and demographics and information on their schools, household structure and parents' education level and income.

Crime

Among other things, CBS has administrative data available on crime suspects, detainees and victims. There are also surveys about the subjective and objective safety of citizens, online safety and crime and the prevalence monitor on domestic and sexual violence.

Health and causes of death

In terms of health, administrative databases are available on hospital admissions, medication use, mental health and health insurance, among others; health surveys can also be used. Causes of death for Dutch citizens are available at the ICD10 level. Health information is available under the same rules that apply to other sources of data.

4.2 Data on Firms

Firms are identified and recorded in units suitable for statistical research in the General Business Register; this is the backbone of the economic statistical process in the Netherlands: other datasets on firms are based on the General Business Register.

Turnover, production and financing

Administrative global data on turnover, production and finance are available for firms, as well as detailed information about turnover, production and costs for large firms and for a sample of small firms. This also includes data on foreign subsidiaries of Dutch companies.

Investments and R&D

For investment, data are available for large firms and for a sample of small firms investment in buildings, land, transport equipment and machinery. Innovation and R&D data are collected on a sample basis. In addition, integral data are available on both patent applicants and applications.

4.3 Environmental Data

Housing

Various aspects about houses are available in the datasets (e.g., location, worth and ownership). There is also a set of files on housing in the Netherlands with preselected variables from registers (varying from house characteristics to household and person characteristics and their income and their wealth). These datasets are intended for research, statistics, the substantiation of plans and the monitoring and evaluation of housing programmes.

Environment

Concerning the environment, there is a dataset available on energy consumption of houses. There is also survey information about environmental costs and environmental investments made by companies and their net environmental costs.

Transport and mobility

In addition to production data on the transport sector, there is also detailed administrative information available about the Dutch motor vehicle fleet and data from a continuous daily survey of residents' travel behaviour.

Distance to facilities

Researchers can use datasets that give the distance measured by road from any object (e.g., a house) to the nearest general practice facility, pharmacy, hospital, public green space (woods, etc.), park, childcare, cinema, museum, supermarket, restaurant, café, etc. In the secure CBS research environment, CBS data users can also calculate the distance between any two objects as the crow flies.

5 Examples of Studies Conducted with CBS Microdata

Some examples might prove helpful in order to give insights into the research possibilities that might be discovered within the secure environment of the CBS.[5] Some projects have a clearly defined scope and research questions. For example, with the introduction of the Work and Social Assistance Act (WWB) in 2004, municipalities became fully financially responsible for the implementation of the welfare assistance for its citizens. Each municipality receives an advance for its welfare duties from the Dutch federal government. This advance is estimated by research institutes SEO and Atlas.[6] This is based on a calculation model that takes the specific population of the

[5] For a full overview, see https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research/institutions-and-projects.

[6] https://www.seo.nl/wp-content/uploads/2023/09/2023-87-SEO-Verdeling-bijstandsbudgetten-2024-achtergronddocument-schattingsfase.pdf.

municipality in question into account and on the basis of their ability to progress from welfare to paid employment using over forty possible indicators that are available in the CBS Microdata environment.[7]

In addition to the provisioning of models, the CBS Microdata facility is also often used for the evaluation of policy. This might include an evaluation of the effects of different COVID support funds for citizens and businesses For example, the Netherlands Bureau for Economic Policy Analysis (CPB) concluded that a special support fund helped self-employed people during the COVID crisis.[8] These projects tend to have a limited and clearly defined scope in terms of research question.

Other researchers have a broader scope or require a multidisciplinary approach. For example, the Amsterdam UMC research on 'Healthy Urban Living' aims to increase an understanding of the environment's contribution to (differences in) health in densely populated areas. They use a multidisciplinary approach, using a combination of health and housing data, but have also added environmental, transport and education data to the project.[9]

Another type of research that frequently occurs within the CBS Microdata facility is the health effect study. Specific sets of one or more population(s) are uploaded by researchers and health data holders and the effects of interventions can be studied.[10] One of the advantages of CBS Microdata is that the population's background features can be mapped, such as in terms of financial background, education, migration background, etc. Researchers must acquire permission from our Expert group on migration statistics in order to use migration background variables. This is undertaken in order to prevent profiling or incorrect conclusions relating to migration background.

6 Pitfalls and Challenges

There are some issues to consider when working with administrative data. Administrations have not been created with research in mind and will, therefore, not always match the researcher's desired application.

Furthermore, administrative data contain errors, changes in the law lead to breaks in the series, as does a change of registration holder or a change in the registration method among registration holders. For example, the Social Support Act (WMO) was introduced in the Netherlands in 2015. This implied that youth care and other types of special care was no longer organised by the federal government, but was instead delegated to municipalities. The administration was no longer handled by one party, but by over 300 municipalities. Many variables were no longer correctly administrated, ranging from empty cells to completely incorrect data.

---

[7] See e.g. Verdeling bijstandsbudgetten 2025 - SEO Economisch Onderzoek (Dutch only).

[8] Een economische analyse van de Tozo regeling | CPB.nl (Dutch only).

[9] https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(23)00258-9/fulltext.

[10] E.g. https://doi.org/10.1016/S2666-5247(24)00079-X and https://bmjopen.bmj.com/content/11/8/e050611.

Researchers working with data on firms have to deal with mergers, separations, closures and restarts of firms, which means that creating a consistent time series is not always easy.

7 Future Outlook

CBS is working hard on the following topics in order to ensure Microdata's relevance for the future:

- Metadata is available in Dutch. Together with ODISSEI (partner of CBS Microdata) a metadata portal is being developed for researchers to access the Microdata metadata in English. This portal will also make it possible for researchers to search on the level of individual variables;
- Due to the continuous growth in use of CBS Microdata and technological advancement, CBS has to keep on par with the technological requests of the users of Microdata. The latest development in this is the linking of the CBS Microdata environment to the ODISSEI Super Computer (OSSC). This virtual IT environment offers a high-performance computer environment that meets the legal, technical and security requirements of Statistics Netherlands. Moreover, the processing of Microdata applications will be made more efficient, researchers will gain greater insight into their projects and the secure Microdata environment will continuously be improved in order to meet the users' ever-evolving needs;
- CBS wants to make the CBS Microdata environment available to as many eligible researchers as possible. At the same time, there are significant costs for users of CBS Microdata as CBS is obliged to bill users for the costs of Microdata research. Our wish is that access to Microdata will be seen as a public good in the future and users will be (partially) reimbursed from public funds.

8 Conclusion

Working with CBS Microdata offers valuable opportunities for researchers and policymakers to generate insights into various social, economic and demographic phenomena. By understanding the structure of the data, conducting thorough analyses and adhering to ethical guidelines, users can harness the power of Microdata to inform decision-making, advance knowledge and drive positive societal change. CBS is pleased to contribute its Microdata to evidence-based policy making, academic research and informed public discourse through the responsible use and dissemination of findings.

Ivo Gorissen graduated in economics from Tilburg University. He has been working at Statistics Netherlands since December 1989. For more than a decade, he has been senior advisor in Microdata Services, dealing with the wonderful and rewarding challenge of combining secure access to data with optimal opportunities for researchers to perform their analyses.

Maarten Bloem works at Statistics Netherlands since 2008, where he did statistical research on Education and Labour and Income. Since 2021 he is one of the senior advisors at Microdata Services.

Vincent van Hees graduated in Human Geography from Utrecht University. He works a little over two year as senior advisor on CBS Microdata. Before that Vincent worked within CBS as a researcher for tailor-made statistics. He knows a little about a lot of CBS Microdata datasets.

Ran van den Boom graduated in Computer Sciences at Leiden University. Since 2022 he is head of Microdata Services at Statistics Netherlands, with almost forty team members. He is also an IPMA-B certified project-and programme manager with experience in the field of data strategy, metadata, administrative processes and communication.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 22
Administrative Data for Research and Public Policy Evaluation in Poland: The Case of Science and Higher Education Sector

Kamila Lewandowska, Ewa Zegler-Poleska, and Emanuel Kulczycki

Abstract This chapter examines administrative data's growing importance for research and public policy evaluation, with a focus on Poland. It provides a dual-level exploration: first, it offers a broad overview of national strategies and key data-sharing mechanisms, and second, it concentrates on sector-specific data in higher education and research. The chapter also includes a user perspective, discussing practical challenges involved in both accessing and utilising this data. While Poland has made notable progress in increasing data availability over the last decade, significant obstacles remain, particularly in effectively processing large datasets and with integrating information from diverse registers.

1 Introduction

Poland was ranked 28th out of 30 countries in the OECD's Open Data Review in 2015 and this reflects a modest performance in terms of public data accessibility and openness. The country has made significant progress since then, now emerging as second in the European Commission's "Open Data Maturity" ranking,[1] which assesses the maturity and progress of European nations in opening public data. This

[1] https://data.europa.eu/en/publications/datastories/open-data-maturity-europe-2023

K. Lewandowska (✉) · E. Zegler-Poleska
Science Studies Laboratory, Centre for European Regional and Local Studies, University of Warsaw, Warsaw, Poland
e-mail: k.lewandowska11@uw.edu.pl

E. Zegler-Poleska
e-mail: e.zegler-poleska@uw.edu.pl

E. Kulczycki
German Centre for Higher Education Research and Science Studies, Berlin, Germany
e-mail: kulczycki@dzhw.eu

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_22

shift has been driven by concerted efforts to both enhance the open data policy framework and to improve the implementation of open data practices.

Despite this commendable progress, Poland continues to face challenges in fully unlocking the potential of open data. While a large volume of public data has been made available, the formats in which they are released often create barriers to efficient data processing and reuse. Issues such as inconsistencies in data formatting and the lack of standardisation across different registers hinder both public and private stakeholders from utilising this data to its full potential.

The chapter is divided into three comprehensive sections. The first section offers a general overview of administrative data in Poland, outlining the national policy frameworks that shape both data collection and its use across various sectors. In the second section, the focus narrows to databases that are specifically relevant to the science and higher education sector. This overview includes a description of the types of data collected and the institutions responsible for data maintenance. The third section shifts to a user perspective, in which the authors share their experiences and findings from utilising administrative data in their own research. This section discusses the practical challenges encountered, such as data accessibility or interoperability issues. The chapter concludes with recommendations for future directions.

2 Public and Administrative Data in Poland: A General Overview

2.1 Legal Framework

In the Polish legal framework, administrative data refers to information collected through official registers and from public administration information systems. These registers encompass data on national economy entities, natural persons, their lives and situation, as well as on phenomena, events, and objects. Public administration information systems, as defined by the Act on Public Statistics of June 29, 1995, collect, gather, and process information managed by authorities, such as the Social Insurance Institution, National Health Fund, and Polish Financial Supervision Authority, among others. Access to public data is primarily governed by the Act of September 6, 2001, on Access to Public Information, which outlines mechanisms by which to make public data available. The Central Public Information Repository (CPIR) offers non-application-based access to public data, thereby removing the need for formal requests, while the Dane.gov.pl portal serves as a central hub through which to access open public data, thereby enabling the reuse of information, particularly in supporting state innovation and the information society. Further legal provisions for the reuse and management of administrative data are outlined in the Act of August 11, 2021, on Open Data and Re-use of Public Sector Information promoting openness and the accessibility of data for diverse purposes.

2.2 Open Data Programme

The Ministry of Digital Affairs launched the Open Data Programme in 2016 to address the challenges presented by accessing public data in Poland—specifically the fragmentation, lack of clear policy objectives, and insufficient strategic direction. This initiative built upon the lessons learned from the danepubliczne.gov.pl portal (currently dane.gov.pl; https://dane.gov.pl/en) that had been established 2 years earlier, in 2014. Key obstacles identified during this period included the excessive use of closed formats such as PDF, poor data preparation quality, a lack of resource descriptions and metadata, the absence of APIs for public databases, and insufficient use of machine-readable formats like XML. Additionally, public institutions often showed reluctance to publish data for reuse, thereby further limiting accessibility and usability.

The Open Data Programme was designed to overcome these challenges by increasing both the quality and quantity of the data made available on the Dane.gov.pl portal. The Programme set out clear standards for sharing public data, drawing on international best practices and insights from the previously deployed portal. The Programme, coordinated by the Ministry of Digital Affairs, involved regular audits to ensure compliance across government institutions. It also incorporated feedback from public consultations and a questionnaire that assessed the societal needs for open data. After being approved by the Council of Ministers, the Programme focused on governmental bodies while also encouraging the wider adoption by both local governments and public institutions.

Poland continues to build on its efforts to open and utilise public data with the adoption of the Open Data Programme 2021–2027, after the expiration of the previous strategy in 2020. It aims to further develop an ecosystem for high-quality data to foster innovation, government transparency, and scientific progress, by fostering an ecosystem of high-quality, accessible data in line with evolving technologies.

2.3 Open Data Portal

Dane.gov.pl serves as a central hub through which to access the data collected by public institutions in Poland. It provides a platform to access a wide array of data from 461 publishers, including 178 public government institutions, 167 local government bodies, and from 116 private entities.

The portal currently offers 42,250 files, 916 websites, and 761 APIs, with CSV being the most common file format, accounting for 24,742 elements. The second most common format is xlsx, with 14,550 elements available. The resources are organised into various categories, offering a comprehensive overview of different sectors. These categories include agriculture, economy, education, energy, environment, government, health, justice, and many others, each with a specific number of resources. For instance, the “Government and public sector” category contains 803

resources, while "Regions and cities" boasts 950; 3621 resources are classified as high-value data.

The portal centralises resources along with metadata, catering to the needs of individuals searching for information for their own purposes. It offers several search options, including by category, data provider, data format, or through a full-text search engine; moreover, it allows users to search for metadata for data providers, data sets, and data in the RDF database using the SPARQL query language. The data catalogue metadata can also be downloaded as a CSV or RDF file. Additionally, advanced users, such as application developers and programmers, can utilise the programming interface (API).

2.4 Data Sharing

Standards for data sharing are detailed in Appendix 1 of the "Public Data Opening Programme."[2] The appendix provides guidelines on data formats for both structured and unstructured data, as well as recommendations concerning the organisation, location, and standardisation of data sources. It also specifies accepted file formats, programming interfaces, and data quality standards, thereby aiming to both streamline and enhance the reusability of public data.

The procedures outlined for data providers include several key stages: the selection of data with high reuse potential, evaluation of its format, adaptation to principles of openness and quality, publication of the data, ongoing supervision of data updates, and assessment of user feedback. The platform supports the automatic generation of certain metadata (e.g., resource identifier, release date, update frequency), while data providers contribute custom metadata (e.g., resource name, description, keywords) to ensure clarity and ease of access for users. The metadata on the portal are predominantly provided in Polish. While some fields—like the resource type and availability date—are generated automatically in a language-neutral manner, the custom metadata (such as resource names, descriptions, and keywords) is supplied by data providers in Polish. The structured approach ensures the quality and consistency of the data shared through the portal, thereby promoting transparency and usability across sectors.

2.5 Data Availability

Data available on the portal are freely accessible for both commercial and non-commercial use without restrictions. Users can download resources from the website directly without the need for logging in or undergoing any verification process, or they can access the data via APIs. Only a small fraction of resources (1806) are

[2] https://dane.gov.pl/en/knowledgebase/useful-materials/program-otwierania-danych-publicznych

listed as protected data, primarily due to concerns about commercial confidentiality, statistical confidentiality, intellectual property rights, and personal data protection. Access to these protected datasets can be requested through the portal under the EU's Data Governance Act. In such cases, public institutions may impose specific conditions or fees for data re-use, ensuring that confidentiality and third-party rights are safeguarded. Data may be made accessible through anonymisation, modification, or secure processing environments, depending on the request made.

While administrative data on the portal is often presented in aggregate form, in order to facilitate public access and policy analysis, the portal also provides extensive access to microdata, which is covered in Sect. 2 in greater detail. A significant volume of microdata is available through platforms such as RAD-on. The platform provides access to granular datasets covering scientific institutions, researchers, and their activities (see Sect. 2). Additionally, while Poland has a vast amount of openly available aggregated administrative data, access to certain types of datasets—such as health records or other sensitive personal data—is subject to legal restrictions due to personal data protection and statistical confidentiality requirements.[3] In such cases, access is possible, but often requires anonymisation or secure processing environments.

The portal serves as the central access point for the data collected by public institutions in Poland. It operates as the "highest-level" system, aggregating and disseminating data from various "lower-level" sectoral systems, such as the national system of information on science and higher education. This hierarchical structure allows users to access a wide range of data from different sectors through a single platform. The following section will focus on the administrative data on science and higher education in Poland in order to provide a more detailed example of administrative data available on the portal.

3 Administrative Data on Science and Higher Education in Poland

Like many other countries (Tomczynska et al. 2023), Poland has been developing a centralised scholarly information system that collects data on scholars and their research, and facilitates the sharing of this information among various services. The need to create such a system has been related to, on the one hand, the necessity to follow global trends and standards related to the open research information—such as FAIR or the recent Barcelona Declaration (https://barcelona-declaration.org/)—and to the increasing role of scientometrics in Poland due to the transformation of the science funding system on the other (Jabłecka & Lepori 2009). In particular, since the first national evaluation exercise of research activity in 1991, the Polish system of research evaluation and funding has evolved from one based on expert assessment to a system that places greater emphasis on scientometric indicators, thereby adopting a more quantitative approach to measuring research productivity (Kulczycki 2017).

[3] https://bip.stat.gov.pl/en/law/statistical-confidentiality/

The demand for extensive data collection has also increased as funding allocation systems grew more complex.

The National Information Processing Institute (OPI PIB) was founded to provide IT infrastructure and support in scientific data collection and analysis. OPI PIB has implemented several database systems, such as the Funding Stream Support System (OSF), to manage grant funding allocation, and the System for Evaluation of Scientific Achievements (SEDN), which enables the Polish Ministry of Higher Education and Science to conduct national assessments of research activity of Polish higher education and research institutions. It is estimated that the institute handles information concerning approximately 11 per cent of Poland's population (Protasiewicz et al. 2021). In 2010, OPI PIB initiated the development of the Information System for Science and Higher Education (POL-on) (Michajłowicz 2021). POL-on collects and stores data provided by Polish universities and research institutions, which are legally required to submit reports electronically. All universities, research institutions, and other science and higher education entities in Poland are required by law to submit reports through POL-on. The system was intentionally designed to standardise the data acquisition process throughout the country (https://polon.nauka.gov.pl).

While some data stored in POL-on is accessible to the public through its user interface, the system does not fully integrate with other open-access scientific databases (Protasiewicz et al. 2021). Consequently, the potential for conducting comprehensive analyses using POL-on is restricted.

3.1 RAD-on

The RAD-on system was developed (Tomczynska et al., 2023) to address this technological gap and to ensure the integration of, and open access to, data on science and higher education in Poland. RAD-on (https://radon.nauka.gov.pl/) provides access to microdata, as well as reports and analyses on higher education and science in Poland obtained from multiple databases administrated by OPI PIB (the system processes 11.47 TB of data) (Protasiewicz et al. 2021). These databases include:

- **POL-on** (https://polon.nauka.gov.pl/): system of information on science and higher education;
- **PBN** (https://pbn.nauka.gov.pl/): Polish Scientific Bibliography, a database with information about Polish scientists' publications;
- **ORPPD** (https://polon.nauka.gov.pl/polon/repozytorium-orpd): Polish National Repository of Theses that stores copies of the theses and dissertations of Polish graduates;
- **Inventorum** (https://inventorum.opi.org.pl/): a platform facilitating the collaboration of scientists with business. It is a recommendation system that provides targeted information about innovation, projects, innovative enterprises, scholars, and scientific institutions;

- **Polish Science** (https://nauka-polska.pl/): a knowledge base on Polish science, including institutions, individuals, scientific work, publications, and conferences;
- **ZSUN/OSF** (https://osf.opi.org.pl/app/): Integrated System of Services for Science/Funding Streams Servicing developed to register and manage applications for research funding competitions organised by the National Science Centre (NCN), the National Centre for Research and Development (NCBR), and the Ministry of Science and Higher Education (MNiSW);
- **SEDN** (https://polon.nauka.gov.pl/sedn2017/pages/login.xhtml): System for Evaluation of Scientific Achievements which handles the process of evaluation of entities that operate in science and higher education in Poland, by using the data obtained from POL-on and PBN;
- **SSSR** (https://recenzenci.opi.org.pl/): Support System for Selection of Reviewers - a knowledge base of potential grant proposal reviewers;
- **ELA** (https://ela.nauka.gov.pl/pl): The Polish Graduate Tracking System.

The RAD-on portal provides access to microdata concerning scientific institutions, activities, academic staff, and promotion procedures and is openly accessible to the public. The platform is equipped with a full-text search engine that facilitates querying the source data derived from the POL-on system and other RAD-on source systems. A full overview of the datasets is available on RAD-on's website in both Polish and in English. Additionally, the metadata accompanying these datasets are also bilingual, so as to support user understanding and navigation of the data.

Microdata on the RAD-on portal are structured at the individual level, containing detailed information about individual scientists, scientific publications, and higher education institutions. Users can explore and scrutinise these records directly on the website, and the data can be viewed, filtered, and downloaded in PDF, XLSX, or CSV file formats. While CSV and XLSX formats provide comprehensive datasets, PDFs offer a more concise version of the information. This multi-format availability ensures that users can access the data in the form that best meets their requirements, whether for detailed analysis or for summary overview.

In addition to microdata, the RAD-on portal provides an analytical module consisting of ready-to-use expert studies and reports prepared by researchers at the OPI PIB, and tools which can be used by users to create their own data summaries. RAD-on resources are freely accessible via a dedicated application programming interface (API), so as to enable users to develop their own solutions and applications that need access to data on higher education. However, in practice, using this API has several limitations (see Sect. 3 on the User Perspective).

3.2 People of Science ("Ludzie Nauki")

While the creation of RAD-on marks significant progress in integrating various data sources, there remains a necessity for a single, reliable repository that consolidates

all information about Polish scientists. Presently, there is no unified source of information on scientists and their achievements; for example, data on academic degrees and employment is available on Polish Science, information on publications can be found on both PBN and RAD-on, and details about doctorates are listed on Polish Science for those awarded prior to 2018 and on RAD-on for those awarded after 2018 (Korytkowski 2023).

Launched in September 2023 by OPI PIB, People of Science ("Ludzie Nauki") aims to serve as the "new CRIS of Poland" (Korytkowski 2023) and aims to establish a central and reliable data source on Polish scientists. The project's first phase is planned to be completed by the end of 2024. The platform will integrate data from multiple existing science and higher education systems, including POL-on, PBN, and Polish Science. People of Science will offer up-to-date, comprehensive, and historical data spanning the last 20 years for approximately 200,000 scientists and doctoral students employed at Polish research institutions. The data will encompass, among other things, scientific degrees obtained, scientific disciplines/fields, employment history at Polish institutions, scientific publications, national and international patents, research projects, doctoral supervisions, dissertation reviews, and external identifiers such as ORCID or WorldCat.

Data on scientific publications will be integrated with external systems such as the Polish National Library, CrossRef, OpenAlex, OpenCitation, and WikiData. Furthermore, People of Science will allow scientists to update and expand their personal information. The data will be accessible via APIs, thereby facilitating the further processing and the creation of new services; the system will also adhere to the FAIR principles (Findable, Accessible, Interoperable, and Reusable) and the Open Government Data concept.

4 Administrative Data on Science and Higher Education: A User Perspective

The authors of this chapter have designed and conducted numerous studies utilising data on the Polish science and higher education sector over the past decade. These studies have primarily focused on four areas:

1. Academic promotion procedures (Kulczycki, 2019);
2. Publication productivity and patterns among Polish scientists compared to their international counterparts using National CRISes (Kulczycki, Engels, et al., 2017; Kulczycki et al., 2018, 2020; Kulczycki & Korytkowski, 2020);
3. Evaluation of scientific activity outcomes at higher education and research institutions (Korytkowski & Kulczycki, 2019a, 2019b; Kulczycki, 2018, 2023; Kulczycki, Korzeń, et al., 2017);
4. Evaluation of artistic activities of higher education institutions in Poland (Lewandowska & Kulczycki 2021; Lewandowska & Kulczycki 2022).

An unusual and significant feature of the studies conducted in the second and third areas was the ability to aggregate outcomes of publication activities and other scientific endeavours, both at the institutional and individual scientist levels. This specificity is notably rare in similar research.

In the analyses related to the second area, the authors utilised data from both the national scientific publication information system (Polish Scientific Bibliography—PBN) and detailed records on individual scientists, including their academic degrees, the years they were awarded, and their current workplaces. For the research in the third area, which focused on evaluating scientific units within the Polish Performance-Based Research Funding system, the data used included activities attributed not only to institutions but also to individual scientists, specifically within their respective scientific disciplines. The research conducted in the first area required advanced data science skills to harvest and to process publicly available data in a manner conducive to analysis.

4.1 First Area: Promotion Procedures

Self-promotion documents prepared by candidates for a habilitation degree (a higher doctorate) from 2011 to 2016 were analysed in the study by Kulczycki (2019). In the promotion process, nearly all documents and reviews prepared by referees were publicly accessible and were required to be posted on the websites of the Degrees and Titles Committee, the national institution overseeing academic promotions in Poland. However, these documents were only available as PDF file links on a Joomla-based website, and the categorisation of documents within the proceedings was neither transparent nor systematic. No search engine was available and the proceedings were not organised in any database.

Additionally, it was not possible to download all documents at once, thereby necessitating the creation of multiple parallel harvesters to retrieve the documents, followed by extensive manual work to associate these files with specific proceedings. Of the 5721 proceedings downloaded, documents were missing in 493 (8.6%) of them, which had to be excluded from the analysis.

4.2 Second and Third Areas: Publication Data and Research Evaluation

The implementation of research in this area was feasible solely due to the privileged position that one of the authors (EK) held at the time as a member of science policy advisory bodies in Poland. This position afforded access to data that enabled the conducting of research and the publication of the study results intended to inform the decision-making processes of both politicians and policy makers.

The authors initially planned to download data on publications by Polish scientists from the PBN system using its API. However, the data could not be downloaded correctly, despite assistance from the API's developers and operators. Consequently, the authors requested a snapshot of the relational database, which was obtained with the support of the Ministry of Science and Higher Education. With this data on scientific publications in hand, information about the Polish scientists who authored these publications was also needed.

Theoretically, all these data are available in the POL-on system. However, practically, it is impossible to download and to link them to the publication data because the necessary identifiers for linking different datasets are not provided.

This issue is also present in the current RAD-on system, which holds a vast collection of data on the Polish science and higher education sector, including information about scientists, their work results, and scientific institutions. The data remains inaccessible for comprehensive analysis because the linking identifiers are unavailable, even though this system also has a functional API. Thus, while users can browse the entire collection of publications by Polish scientists and can search for publications by titles, it is impossible to download a list of publications by Polish scientists in a specific discipline, due to the incomplete data linkage.

4.3 Fourth Area: Evaluation of Artistic Activities of HEIs

In the studies by Lewandowska & Kulczycki (2021, 2022), the authors analysed the evaluation of artistic activities of higher education institutions within the Polish Performance-Based Research Funding System. The data analysed concerned 136,894 outputs submitted by art faculties for two evaluation exercises (in 2013 and 2017). The data used was retrieved from two sources: the POL-on system and the SEDN (System for Evaluation of Scientific Achievements).

While data on artistic achievements submitted to the national research evaluation system is publicly available in the POL-on database, and is accessible via the RAD-on system, there are several limitations related to the use of this data. Firstly, artistic outputs are non-traditional (non-written) research outputs that can take many different forms, thereby resulting in a lack of standardised categorisation. This contrasts with traditional research outputs that typically take the form of journal articles, books, or chapters. In the Polish research evaluation system, the categorisation of artistic outputs and the types of outputs eligible for evaluation have varied from one evaluation exercise to the next. This inconsistency makes it impossible to compare different evaluation exercises and to evaluate the effects of science policy accurately. Secondly, the data on artistic outputs includes only short text descriptions and no non-written material (such as visual files, audio, or video). This absence of supplementary material makes it difficult to verify the data against the results of the evaluation process, thereby undermining the transparency of research assessment. Finally, it is impossible to link the data on artistic outputs to other datasets (e.g.,

publication data) because the necessary identifiers for linking different datasets are not provided.

Final Considerations on Data Access and Use

Poland has made notable progress in developing policies that promote an open data culture over the past decade. The creation of the Open Data Programme and the Open Data Portal, which include standardised procedures for data sharing required of public institutions, has significantly advanced the country's open data landscape. Additionally, efforts have been made to enhance sector-specific systems, such as those in science and higher education. Platforms like danepubliczne.gov.pl and RAD-on offer a substantial amount of administrative data, including microdata, which users can access directly via the websites without needing to preregister, log in, or to apply for access. There are no specific requirements (such as an affiliation with a Polish institution, etc.) to access this data.

Despite these advancements, challenges persist in effectively using and processing the large volumes of data, particularly when attempting to integrate information from different registers. One of the main issues concerns the lack of standardised unit-level identifiers that allow for tracking individuals over time and across datasets. For example, linking individual researchers to their publications, institutions, and assigned disciplines is often difficult without additional identifiers in the science and higher education sector. These identifiers are unavailable due to system design limitations. Improving the availability and structure of these identifiers is critical for better data integration and for more comprehensive analysis. The absence of linking keys between datasets limits the possibility to conduct multi-dimensional research, thereby reducing the depth of insights that can be drawn from the data.

Our analysis focuses specifically on data from the higher education and research sectors. While this focus has allowed us to provide an in-depth examination of administrative data accessibility in these areas, we recognise that data access issues may differ in other sectors. However, we believe that our observations capture key trends and challenges that are representative of the national open data environment, given the broader policy landscape and general data-sharing mechanisms in Poland. While we cannot fully assess external validity across all fields, our findings provide insights into systemic patterns of data accessibility, usability, and integration that likely extend beyond the sectors that we have examined.

As Rogalińska (2022) highlights, integrating resources is crucial for effectively utilising information from administrative registers. Establishing legal and organisational frameworks that support this integration is vital for further opening administrative data. An example is the Ministry of Science and Higher Education's development of a new system called "Ludzie Nauki" (People of Science), which is designed to aggregate data about individual scientists. However, as of May 2024, only a basic version of this system—focused on researchers' publications, workplaces, and academic promotions—is available, thereby making it difficult to assess its potential impact.

Reflecting on our experiences, we recommend the following best practices for enhancing open data access while maintaining privacy compliance. Systems

collecting and sharing data should facilitate data aggregation across different registers. Clear guidelines and procedures for data access, including public keys for data linkage, should be established to streamline and secure the process. Additionally, systems should be designed to allow for easy data integration and linkage across different sources and formats, thereby ensuring seamless connectivity and comprehensive analysis. Implementing these practices will improve the utility of open data while safeguarding privacy.

References

Act of June 29, 1995 on Public Statistics [Ustawa o statystyce publicznej z 29 czerwca 1995 r.]. Journal of Laws [Dziennik Ustaw] of 1997 no. 88 item 439. https://isap.sejm.gov. pl/isap.nsf/DocDetails.xsp?id=wdu19950880439.

Jabłecka, J., & Lepori, B. (2009). Between historical heritage and policy learning: The reform of public research funding systems in Poland, 1989–2007. *Science and Public Policy, 36*(9), 697–708. https://doi.org/10.3152/030234209X475263

Korytkowski, P. (2023). Ludzie Nauki: A new CRIS for Poland. *Accessible at*. https://dspacecris.eurocris.org/handle/11366/2505

Korytkowski, P., & Kulczycki, E. (2019a). Examining how country-level science policy shapes publication patterns: The case of Poland. *Scientometrics, 119*(3), 1519–1543. https://doi.org/10.1007/s11192-019-03092-1

Korytkowski, P., & Kulczycki, E. (2019b). Publication counting methods for a national research evaluation exercise. *Journal of Informetrics, 13*(3), 804–816. https://doi.org/10.1016/j.joi.2019.07.001

Kulczycki, E. (2017). Assessing publications through a bibliometric indicator: The case of comprehensive evaluation of scientific units in Poland. *Research Evaluation, 26*(1), 41–52. https://doi.org/10.1093/reseval/rvw023

Kulczycki, E. (2018). The diversity of monographs: Changing landscape of book evaluation in Poland. *Aslib Journal of Information Management, 70*(6), 608–622. https://doi.org/10.1108/AJIM-03-2018-0062

Kulczycki, E. (2019). Field patterns of scientometric indicators use for presenting research portfolio for assessment. *Research Evaluation, 28*(2), 169–181. https://doi.org/10.1093/reseval/rvy043

Kulczycki, E. (2023). *The evaluation game: How publication metrics shape scholarly communication*. Cambridge University Press.

Kulczycki, E., Engels, T. C. E., & Nowotniak, R. (2017). Publication patterns in the social sciences and humanities in Flanders and Poland. In *16th international conference on Scientometrics and Informetrics (ISSI 2017)* (pp. 95–104).

Kulczycki, E., Engels, T. C. E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., Nowotniak, R., Petr, M., Sivertsen, G., Istenič Starčič, A., & Zuccala, A. (2018). Publication patterns in the social sciences and humanities: Evidence from eight European countries. *Scientometrics, 116*(1), 463–486. https://doi.org/10.1007/s11192-018-2711-0

Kulczycki, E., Guns, R., Pölönen, J., Engels, T. C. E., Rozkosz, E. A., Zuccala, A. A., Bruun, K., Eskola, O., Starčič, A. I., Petr, M., & Sivertsen, G. (2020). Multilingual publishing in the social sciences and humanities: A seven-country European study. *Journal of the Association for Information Science and Technology, 71*(11), 1371–1385. https://doi.org/10.1002/asi.24336

Kulczycki, E., & Korytkowski, P. (2020). Researchers publishing monographs are more productive and more local-oriented. *Scientometrics, 125*(1), 1371–1387. https://doi.org/10.1007/s11192-020-03376-x

Kulczycki, E., Korzeń, M., & Korytkowski, P. (2017). Toward an excellence-based research funding system: Evidence from Poland. *Journal of Informetrics, 11*(1), 282–298. https://doi.org/10.1016/j.joi.2017.01.001

Lewandowska, K., & Kulczycki, E. (2021). Academic research evaluation in artistic disciplines: The case of Poland. *Assessment & Evaluation in Higher Education, 47*(2), 284–296. https://doi.org/10.1080/02602938.2021.1893651

Lewandowska, K., & Kulczycki, E. (2022). Academic research evaluation in artistic disciplines: The case of Poland. *Assessment & Evaluation in Higher Education, 47*(2), 284–296. https://doi.org/10.1080/02602938.2021.1893651

Michajłowicz, M. (Ed.). (2021). *I Systemy informatyczne wspierające naukę i szkolnictwo wyższe POL-on: System Informacji o Nauce i Szkolnictwie Wyższym TOM I*. National Information Processing Institute. Accessible at: https://opi.org.pl/wp-content/uploads/2021/10/POL-on.TomI-.pdf

Protasiewicz, J., Podwysocki, E., Ostrowska, S., & Tomczyńska, A. (2021). Open access to data on higher education and science: A case study of the RAD-on platform in Poland. In 2021 EPiC series in computing (Vol. 78, pp. 9-21). https://easychair.org/publications/paper/6pSZ

Rogalińska, D. (2022). Wtórne wykorzystanie danych administracyjnych w statystyce publicznej—obecnie i w przyszłości. *Accessible at*. https://www.parp.gov.pl/component/publications/publication/wtorne-wykorzystanie-danych-administracyjnych-w-statystyce-publicznej-obecnie-i-w-przyszlosci

Kamila Lewandowska is Assistant Professor at the Science Studies Lab, EUROREG, University of Warsaw. She is currently coordinating an international grant on practice-based research, co-funded by the National Science Centre (NCN), the Swiss National Science Foundation (SNSF), and the Research Foundation—Flanders (FWO). Her research centers on innovative strategies for integrating research with practical applications.

Ewa Zegler-Poleska is Research Assistant at the Science Studies Lab hosted by EUROREG, University of Warsaw. She is completing her PhD in Information Science at Indiana University Bloomington. Her interests include research data as a form of scholarly output.

Emanuel Kulczycki is the head of Scholarly Communication Research Group (Adam Mickiewicz University in Poznań) and a policy advisor to the Ministry of Higher Education and Science. The former chair of the European Network for Research Evaluation in the Social Sciences and the Humanities and a co-founder of the Helsinki Initiative on Multilingualism in Scholarly Communication.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 23
Administrative Data in Portugal

Francisco Lima, José Pinto Martins, Katja Neugebauer, Joana F. Pimentel, Sofia Rodrigues, Paulo Saraiva, Rita Sousa, and João Zilhão

Abstract This chapter examines the administrative data access policies in Portugal and focuses on two key institutions: Statistics Portugal and Banco de Portugal. It explores their respective approaches to balancing researcher access with data confidentiality. The chapter's first part analyses Statistics Portugal's integration of administrative data into official statistics, highlighting the enhanced analytical capabilities provided by the National Data Infrastructure. It also identifies persistent challenges, particularly in terms of timely data provision. The second part describes data access through Banco de Portugal's Microdata Research Laboratory (BPLIM), detailing the legal framework, access procedures, and key datasets that are available to researchers. The varying degrees of data anonymisation employed are discussed, illustrating the trade-offs between research utility and the protection of privacy.

F. Lima (✉)
University of Lisbon, Lisbon, Portugal
e-mail: francisco.lima@tecnico.ulisboa.pt

J. P. Martins · S. Rodrigues · P. Saraiva · J. Zilhão
National Statistical Office INE, Lisbon, Portugal
e-mail: pinto.martins@ine.pt

S. Rodrigues
e-mail: sofia.rodrigues@ine.pt

P. Saraiva
e-mail: paulo.saraiva@ine.pt

J. Zilhão
e-mail: mjoao.zilhao@ine.pt

K. Neugebauer · J. F. Pimentel · R. Sousa
Bank of Portugal, Porto, Portugal
e-mail: kneugebauer@bportugal.pt

J. F. Pimentel
e-mail: jmpimentel@bportugal.pt

R. Sousa
e-mail: rcsousa@bportugal.pt

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_23

1 Institutional Background

The National Statistical Law of 2008[1] defines the general framework of the National Statistical System (NSS) in Portugal. The NSS consists of the following entities and services: the Statistical Council, a governmental body tasked with overseeing and coordinating the system; Statistics Portugal, responsible for both generating and disseminating official statistics, as well as ensuring the supervision and technical and scientific coordination of the NSS; Banco de Portugal, which, as part of its mission, collects and compiles statistics related to monetary, financial, foreign exchange, and balance of payments matters; the Regional Services of Statistics of the Autonomous Regions of Azores and Madeira, which serve as extensions of Statistics Portugal for national statistics and as regional offices for regional statistics; and other entities authorised by Statistics Portugal to produce official statistics.

Access to administrative data collected by Statistical Authorities is possible at the Research Data Centres (RDCs) of Statistics Portugal[2] and Banco de Portugal.[3] Additionally, the Ministry of Education and Science[4] has a dedicated safe centre to facilitate access to Education, Higher Education, and Science data for research purposes.

2 Access to Administrative Data Through Statistics Portugal

Statistics Portugal envisages having a stronger role in the context of a rapidly changing data ecosystem. It is, therefore, imperative to embrace new technologies, new methodologies, and new sources in order to stay relevant and to provide accurate, timely and relevant information. In practical terms, Statistics Portugal needs to do more with less.

Statistics Portugal has been actively engaged in an innovation strategy, developing and strengthening, among others, its data management and analytical capabilities in order to better serve the needs of both the country and the international community.

This section offers a comprehensive overview of how the organisation has adapted to these new needs, taking the increasing availability and complexity of non-statistical data sources into consideration, particularly administrative data, through the implementation of the National Data Infrastructure.

[1] Lei do Sistema Estatístico Nacional, Law n. 22/2008 of May 13.

[2] www.ine.pt.

[3] www.bportugal.pt.

[4] www.dgeec.medu.pt.

2.1 The National Data Infrastructure: A Holistic Approach

The increasing and intensive use of non-statistical data sources (administrative and others) brings important challenges to the official statistics production process. Statistics Portugal has developed several initiatives aiming to respond to these challenges in recent years in which the creation of the National Data Infrastructure is highlighted as necessary.

National Data Infrastructure is a system developed by Statistics Portugal to both integrate and utilise administrative data and other sources for official statistics. It provides a single access point to various data types through various tools and for multiple users and purposes. The system ensures data protection and integrity throughout the production chain, from platform development and data collection to validation and analysis, adopting both a more intensive and more integrated approach to producing statistical information (Fig. 23.1).

- Use multiple data sources.
- Increase the analytical capacity and flexibility in data management.
- Exploit new data sources for statistical and research purposes.
- Foster innovation.

National Data Infrastructure embodies several strategies and initiatives that Statistics Portugal has implemented in recent years. These efforts are important on their own, certainly, but they are much more impactful when working together, given that they were designed to create a comprehensive and integrated data ecosystem that supports the needs of various stakeholders, including government agencies, businesses, researchers, and the general public. At an **organisational** level, Statistics Portugal adjusted its internal organisation, revised the Statistical Production Handbook, and established a dedicated Administrative Data Unit.

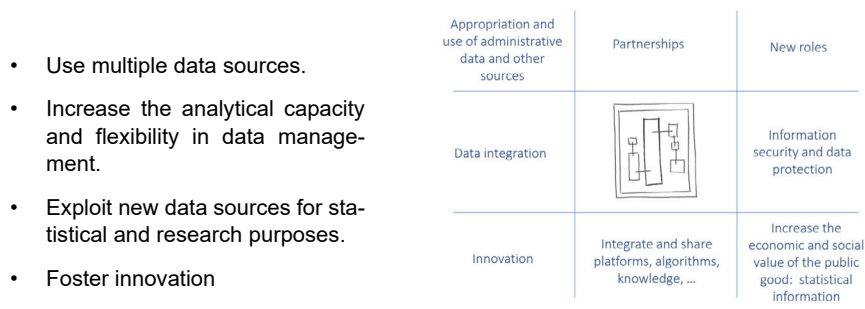


Fig. 23.1 National Data Infrastructure and its main purposes

At a **technological** and **methodological** level, Statistics Portugal has significantly invested in IT infrastructure, data analysis tools, and employee training in order to enhance data exploration and analysis by using artificial intelligence and machine learning. Data protection and confidentiality are paramount to Statistics Portugal, which maintains a vigilant approach, recognising the ever-evolving landscape of data protection techniques. Data from external sources, identified by unique identifiers (e.g., tax number, citizen ID), are encrypted using the irreversible SHA-256 algorithm prior to transmission to Statistics Portugal. Specific datasets, subject to stringent access controls, isolate sensitive personal data (name, address, etc.). All data access is rigorously logged for audit purposes.

At a **dissemination** and **communication** level, Statistics Portugal has implemented new strategies to acquire data from data holders, including direct engagement and multidisciplinary negotiations. It has maintained strong partnerships with data providers and is developing feedback reports in order to enhance collaboration. Additionally, it has provided tools and resources for internal users to support daily work with administrative data. From the users' perspective, researchers can access high-quality data through anonymised datasets and pre-processed tables, with detailed metadata and a strong investment in the availability of administrative microdata for research. Statistics Portugal makes statistics accessible and understandable for the general public through clear and concise publications, press releases, and media outreach. It also provides information about how statistics are produced to build trust and transparency.

Having established the framework in which Statistics Portugal deals with administrative sources and their potential, we will now focus on practical applications. The sections that follow will delve into concrete examples of best practices surrounding administrative data sources, identifying their main strengths and limitations, and some examples of use cases.

2.2 Simplified Business Information

The Simplified Business Information was created in 2006 within the framework of a government programme for simplifying and modernising Public Administration, named the SIMPLEX programme.

This measure, which resulted from a joint effort of all of the public administration entities involved (the Ministry of Finance, the Ministry of Justice, Statistics Portugal and Banco de Portugal), aggregates the fulfilment of four previously dispersed legal obligations by enterprises in a single act. With the Simplified Business Information, the enterprises' obligations, namely the settlement of annual accounts, are fully complied with electronically and are carried out on a single occasion.

This measure significantly impacted both enterprises—around 450,000—and the different entities of the Public Administration responsible for collecting this information.

The subsequent steps of the simplification process comprise pre-filling the enterprise's declaration with data extracted from the Standard Audit File for Tax Purposes (SAF-T), eliminating variables and tables from current forms.

The main advantages of Simplified Business Information include complete coverage of the business universe—around 450,000 compared to the 50,000 enterprises surveyed previously; the reduction of information availability from 12 to 6.5 months; information received automatically by electronic means; a significant increase in the detail of the information; and the elimination of one of the costliest surveys carried out by Statistics Portugal. Additionally, sectoral surveys (Industry, Construction, Retail trade, Services, etc.) also benefit from Simplified Business Information data in both the sample design and statistical production, thereby impacting quality while ensuring consistency with the Portuguese business framework. The Simplified Business Information microdata, integrated at the enterprise level with other administrative and survey data, increases the possibilities for statistical production and research (without imposing additional reporting burdens on enterprises). Examples of microdata being integrated with Simplified Business Information include administrative data, such as e-Invoice (see below) and VAT from Tax Authority, Monthly Statement of Remunerations from Social Security, and data from Community Innovation Survey, International Trade in Goods statistics, and the Information and Communication Technologies Usage Survey.

While a valuable tool for streamlining business reporting, Simplified Business Information is not without its challenges, though. Overall, the most significant limitations are as follows:

- Data must be submitted by July 15th of the year following the reporting period, but, the Government has unilaterally postponed this deadline on several occasions without prior approval, thereby significantly impacting the production of those statistics entirely based on the Simplified Business Information;
- Regulatory changes that already occurred in the accounting framework, namely the implementation of new/revised International Financial Reporting Standards (IFRS), are not yet implemented in Simplified Business Information;
- The evolution of the system, namely the pre-filling and the detailing of some data, has not yet been implemented, due to resistance from some sectors, specifically the Order of Chartered Accountants, which considers the change proposed to be excessively demanding and detailed;
- Implementing necessary changes to the Information System is a lengthy process, one hindered by bureaucratic hurdles such as decree approvals and technological developments. Moreover, businesses face significant disruptions as they must update their software in order to accommodate these changes.

2.3 The e-Invoice

The e-Invoice system is an electronic invoicing software system adopted by the Portuguese government in January 2013 to fight VAT fraud.[5] Since 2013, the law requires the electronic reporting of invoices to the Tax Authority and covers all individuals or legal entities with a headquarters, stable establishment, or tax domicile in Portugal. The data generated by e-Invoice cover all business-to-business and business-to-consumer transactions.

Since May 2020, Statistics Portugal has received monthly data—roughly 100 million records per month—on business transactions and personal expenditures (net of indirect taxes) declared in electronic invoices per a protocol set up between the Tax Authority and Statistics Portugal. The e-Invoice data treatment has played an essential role as proof of the success of Statistics Portugal's administrative data strategy. The procedures created—e.g., metadata produced and service level agreements for internal and external users—are now being applied to all other administrative sources. (Part of) e-Invoice microdata has been available to the research community since December 2023, alongside a characterisation report, which identifies data structure, definitions, and the most relevant procedures applied. Regular feedback to data providers is under development.

Strengths include the e-Invoice's unique characteristic of having information from both seller and buyer, thereby generating many opportunities and challenges to both exploring and integrating with other data. From the buyer's perspective, one can study and produce new statistics on consumer behaviour. From the seller's perspective, one main goal is to replace monthly turnover in all enterprises' surveys (Short-Term Statistics). Original data was also improved and enriched with structural data from Statistics Portugal (namely activity code, region, etc.) and can be integrated with all other enterprise-level microdata by researchers. The next step will be making individual-level integration (the buyer's perspective) available for research. Conversely, limitations may be described as follows:

- Due to Statistics Portugal's needs to have timely data, original e-Invoice data raw format needs extensive cleaning and standardisation processes, which can be time-consuming and resource-intensive as the volume of data is quite high (+100 million records per reference month);
- Reporting delays for e-Invoice to the Tax Authority, especially from large enterprises, hinders the quality of timely data availability for statistical analysis;
- The Tax authority does not always meet the deadlines for the transmission of e-invoice information to Statistics Portugal, primarily due to IT-related issues such as the need to update servers. This can hinder the goal of replacing the current monthly short-term statistics surveys.

[5] Portal das Finanças (portaldasfinancas.gov.pt)

2.4 The DNS

The data provided by DNS.PT refers to data concerning the registration and maintenance of internet domains that, in turn, can support one or more websites/portals. In a simplified way, a domain name identifies the location of a site on the internet and consists of a unique name followed by a dot and a suffix, such as ".pt" for Portuguese domains. DNS.pt data includes details on the domain's status (active, inactive, or expired), registration and expiration dates, and the technical and administrative contacts for each domain.

This comprehensive dataset is very useful to track the growth of online presences in Portugal and provides insights into the country's trends in web developments. By integrating this information, Statistics Portugal can better analyse the digital landscape and its impact on various sectors. One of the most interesting features of this source is the interaction between the data provider and Statistics Portugal. The data owner is very interested in this joint work, and the dataset transmitted to Statistics Portugal has been improved twice. A dashboard, jointly produced by Statistics Portugal and DNS.PT, was released in the third quarter of 2024. Conversely, each domain has a technical manager, an administrative manager, and an owner and this information is not always correctly defined, which can cause some bias in the association of domains with their real owner and might either hinder or postpone a more intensive use of this data, namely in the integration with enterprise information and knowledge society data.

2.5 Census with Administrative Data

Population and housing censuses represent one of the pillars of any National Statistical System concerning the enumeration and characterisation of the population and households at the national, regional, and local levels, thereby producing essential indicators for the definition of public policies and decision-making. Statistics Portugal is working on the transformation of the traditional census model into a more efficient one by using administrative data.[6] This is Statistical Portugal's most ambitious project of data integration from administrative sources due to its dimensions.

Within the scope of the Census, Statistics Portugal has been studying the contribution of available administrative data with the potential to replace information collected through full enumeration (the traditional model) after the 2021 Census. The project is a key element in the paradigm-shifting strategy and places Portugal in a more favourable position in the international context, namely by being able to produce annual census statistics.

[6] Census with administrative data

Central to the project is the constitution of the Resident Population Database (the statistical population registers) that covers a set of characteristics—geographical, demographic, and socio-economic—of the resident population in Portugal and results from the integration of information from various public administration sources. Administrative sources integrated in Resident Population Database include: Civil Register; Tax register; Social Security Register; State Pension and Work Fund Register; Unemployment Register; Immigration register; Education Register; and the Private Employment Register. Statistics Portugal establishes partnerships with data providers through written agreements specifying the registration design, transmission dates, and information on security measures, complying with both the provisions of the National Statistical System law and the General Data Protection Regulation.

Portugal's administrative census is a valuable tool for understanding the country's population, economy, and social conditions and its major strengths are related to:

- Cost-effectiveness: by leveraging existing administrative data, Portugal's census significantly reduces the need for expensive fieldwork and data collection operations, thereby making it a more financially sustainable approach;
- Efficiency: the reliance on pre-existing records streamlines the data collection process, allowing for quicker and more efficient analysis;
- Timeliness: administrative data is often more up to date than traditional census data, thereby providing policymakers and researchers with timely insights into the country's population, economy, and social conditions.

Furthermore, the establishment of a Resident Population Database is anticipated to significantly enhance both the accuracy and representativeness of sampling frames. By focusing on individuals, rather than on households, this database will enable more precise targeting of specific population segments for surveys and studies. Limitations are related to:

- Coverage, as the census with administrative data may not capture certain population segments accurately, such as undocumented immigrants, homeless individuals, or people who have not interacted with government agencies;
- Data quality, as with administrative data it can vary depending on the specific data provider and on the type of data collected;
- Privacy concerns, because using administrative data with this kind of sensitivity raises concerns about privacy and data protection;
- Limited scope, as the census with administrative data may not provide as detailed information as would be found in a traditional census, especially for certain demographics or specific topics.

2.6 Data Provision and Access for Research Purposes

According to the Law on the National Statistical System (Law No. 22/2008), individual data on persons, households, enterprises, and organisations can be provided for scientific purposes in an anonymised form.[7] Statistics Portugal provides free access for researchers to the information (both from administrative sources and surveys) needed to carry out their activities.

Statistics Portugal, by principle, can provide access to all surveys used in the production of official statistics. Additionally, access to administrative data may be granted with the authorisation of the data holders. In 2019, a significant milestone was reached with the signing of a protocol between Statistics Portugal and the Tax Authority. This agreement enables Statistics Portugal to make anonymised microdata from the Tax Authority accessible to the scientific community, theoretically opening the door for all administrative data flows to become available. However, sufficient time must be allocated to both understand and prepare the data (including metadata) and to conduct output checks.

The data is accessible with all enhancements made by Statistics Portugal and accompanying metadata. Access is exclusively granted through the secure data centre. Statistics Portugal is working to include similar authorisation provisions in protocols with other data providers. In October 2024, 72 (seventy-two) microdatabases were available for scientific research purposes, with the following thematic distribution:

Topic	No. of Databases
Living conditions of families	9
Companies	14
International trade	1
Construction	8
Demography	6
Labour market	5
Agriculture	2
Industry and energy	2
Information society	4
Health	5
Education and training	2
Culture	2
Science and technology	4
Economic situation	4
Transport and communications	1

(continued)

[7] As established equivalently in the EU Regulations 223/2009 on European statistics and 557/2013 on access to confidential data for scientific purposes.

(continued)

Topic	No. of Databases
Environment	1
Tourism	1
Justice	1
Total	**72**

The complete list of micro-databases for scientific research, with the periods available and the corresponding metadata that can be accessed, is available on the respective website at Portal do INE.

Researchers must be affiliated with entities dedicated to research, such as R&D institutions, universities, or international organisations in order to get access to data. Researchers must obtain accreditation from Statistics Portugal, which is based on their affiliation and the nature of their research, namely:

- Researchers working in entities, structures, and networks dedicated to the production, dissemination, and transmission of knowledge in the national science and technology system (R&D institutions, collaborative laboratories, technological interface centres, science and technology infrastructures, networks and consortia of science and technology) and organisations, institutions, or departments certified as targeting research purposes;
- Students with ongoing doctoral theses and registered in the official national register of doctoral theses;
- Master's degree students, provided that the request is signed off by a supervisor of the thesis in progress, a professor at a legally recognised national higher education institution;
- Researchers with proven affiliation in international organisations, such as specialised agencies/structures of the European Commission; specialised agencies of the United Nations (International Labour Organisation (ILO), Food and Agriculture Organisation (FAO), United Nations Educational, Scientific and Cultural Organisation (UNESCO), World Health Organisation (WHO), World Bank Group; International Monetary Fund (IMF); and Organisation for Economic Cooperation and Development (OECD);
- Researchers from other research organisations that appear on the list of entities recognised by Eurostat on the date of data access, pursuant to Article 4 of Regulation no. 557/2013 implementing Regulation no. 223/2009 of the European Parliament and of the Council on European statistics, concerning access to confidential data for scientific purposes.

The accreditation is valid during the declared length of the research project and only for the data identified in the request. It requires the signature of a Code of Conduct by the applicant and the research institution of affiliation. The researcher applies for a research project, naming all researchers involved and specifying the

project's duration, objectives, and methods. Each researcher must sign a form and a declaration of confidentiality. The description of the procedure and the relevant supporting documents are available on the Statistics Portugal website under Researchers.

Access to microdata is free of charge. Three access modes are possible according to the type of microdata:

- Fully anonymised data files—scientific use files—made available by a secure Statistics Portugal's cloud (two-factor authentication). Data has been pre-processed to eliminate the need for output checking and researchers can use the files in their place of work;
- Ready-made tables that are also made available by cloud (two-factor authentication);
- Data files—secure use files—available only in a Safe Centre environment and subject to output check.

In addition, Statistics Portugal has prepared some files, with data treated, so that the observation unit cannot be identified directly or indirectly—the so-called Public Use Files (PUFs). They are directly available for download from Statistics Portugal's website and imply prior acceptance of the conditions of use.

2.7 Data Integration and Data Analysis

All microdata is treated and structured so as to be research-ready, accompanied by the respective metadata. Statistics Portugal's website provides the list of datasets and metadata (although not all are in English).

All observation units in the microdata have an (anonymous) unique identifier that allows for longitudinal analysis when applicable. At the enterprise level, the unique identifier enables the integration of different datasets from survey and administrative sources, thereby raising the dimensions of analysis for researchers.

The standard software installed on the computer is SPSS, R, and *Stata* (licensed by Statistics Portugal). Researchers are free to run their own code within the available software. Researchers can also incorporate external information into their analysis, but it must be clearly documented and justified within the research project.

3 Access to Administrative Data from Banco de Portugal

Alongside national statistical offices, central banks are also particularly well-positioned to bridge the gap between producers of administrative microdata and the research community. This is because they either collect and produce microdata sets or have privileged access to administrative datasets, while at the same time having a body of in-house researchers that routinely use the data for their research activities.

Banco de Portugal has accumulated extensive expertise in the collection, handling, and production of microdata sets over the years. Additionally, several researchers at the Bank have successfully built research agendas centred on the exploration of Portuguese administrative micro-databases. Thus, the goal of setting up a research unit dedicated to the analysis of microdata was to harness the power of the Portuguese administrative micro datasets, thereby further stimulating research anchored on these data.

Access to microdata from Banco de Portugal is provided by the Banco de Portugal Microdata Research Laboratory (BPLIM). The focus of the following section will, therefore, be on BPLIM and how it provides researchers access to administrative data.

3.1 About Banco de Portugal Microdata Research Laboratory (BPLIM)

Banco de Portugal collects a vast amount of financial and economic microdata as a regulator and a national statistical authority. While the collection's primary goal is the production of official statistics or the verification of compliance with regulations, the information collected is also important for the development of evidence-based policies and insights that can positively impact the Portuguese economy and society. It was in line with this initiative that Banco de Portugal established BPLIM in 2016, with the main objective of facilitating external research access to the microdata sets collected and maintained by the Bank. BPLIM is based at the Porto office of Banco de Portugal and is a unit of the Economics and Research Department of Banco de Portugal. It is staffed by a team that includes data analysts and researchers with extensive expertise in microdata.

BPLIM operates with several key objectives in mind. Firstly, it was designed with both the national and international research communities in mind. Its website,[8] along with all documentation, is produced in English. BPLIM allows for different forms of non-presential access (remote and surrogate access, as detailed in the following section). Second, it implements rigorous measures for both confidentiality and data security to ensure the integrity and privacy of microdata. BPLIM microdata never leaves the infrastructure of the Bank and is always analysed on its dedicated servers. Finally, BPLIM pays particular attention to the issue of reproducibility in order to help strengthen the credibility of research findings and to accommodate a changing scientific environment.

BPLIM invests many resources in the creation of efficient, consistent, and well-documented datasets that can be readily used for research purposes. It also develops templates and data reports that facilitate the analysis of those datasets. This alleviates

[8] https://bplim.bportugal.pt.

the necessity for the researcher to do substantial "data-cleaning" and preparation tasks and provides consistency of analysis across research projects and researchers. BPLIM encourages an open dialogue with researchers to understand their exact requirements. Furthermore, it strives to answer their data needs, providing them with solutions that do not compromise the confidentiality of the data. BPLIM also invests many resources in the creation of tools for recurrent data tasks[9] as well as other tools that can help researchers more generally. All of these are made available in BPLIM's GitHub repository.[10] Finally, BPLIM staff provides assistance in order to tackle and resolve any potential issues related to scientific and computational support, statistical software usage, and microdata exploration.

Another important aspect of BPLIM's initiatives is its commitment to enhancing the research community's technical expertise. Every year, the laboratory offers assistantships and internships for young researchers. BPLIM has already welcomed around 40 master's students since the beginning of its activity. Additionally, it hosts complimentary workshops designed to bolster researchers' skills in various areas. These workshops feature leading experts who share their insights, best practices, and the latest developments in data analysis.[11]

BPLIM has been consistently expanding its activity, in terms of the number of projects and researchers. Over 250 projects were approved from 2016 until the end of 2024. These projects cover a variety of topics, focusing on economic and financial topics.[12] At the end of 2024, there were around 130 active projects, involving more than 150 external researchers, representing 14 Portuguese universities, 32 foreign universities (predominantly from Europe and the USA), and 14 non-university research institutions, including international organisations. If one considers all of the projects approved since its inception until the end of 2024, BPLIM has already involved close to 350 researchers, mostly external (83%) and from national institutions (66%). This shows that the fraction of researchers affiliated with foreign institutions is already considerable.

3.2 Data Access

Researchers can access comprehensive information about how to submit a project and access data on the BPLIM website. This information includes both general and specific documentation (e.g., *Application Form*, *Confidentiality Agreement*, *Guide for Researchers*, etc.). Specific documentation about individual datasets, such as *Manuals* and metadata, is also easily accessible in English.

[9] For example, through the construction of a panel or of specific variables.

[10] BPLIM's Github page: https://github.com/BPLIM

[11] https://github.com/BPLIM/Workshops

[12] For a complete list of all projects, see https://bplim.bportugal.pt/projects

Data access is only allowed for researchers who intend to use it for scientific purposes. BPLIM does not distinguish between domestic and foreign nationals, with respect to data access, and no affiliation with a Portuguese institution is required. Researchers can submit their projects at any time, with the acceptance of proposals based solely on their scientific merit. However, server capacity and staff availability may also impose certain limitations. There are no additional restrictions beyond these factors.

Internal and external researchers are requested to complete an *Application Form,* providing a comprehensive description of the project, including the identification of all researchers and their affiliations, the datasets, the timeframe, and the variables. BPLIM staff will work with the researcher(s) to identify the required datasets and, if needed, to build customised dataset(s). For external researchers, the submission of a *Curriculum Vitae* is mandatory, alongside the signing of *a Confidentiality Agreement* by all researchers involved in the project. Moreover, in cases where the project is conducted within the preparation of a Master's or PhD thesis, the active participation of the respective supervisor is imperative. Furthermore, external researchers are required to be affiliated with a research institution and have no conflict of interest, and their projects must demonstrate scientific merit.

A decision is taken after an external researcher submits the project proposal, typically, within a week. If the research project is approved, an account is opened on the BPLIM external server, where the data can be accessed remotely using a safe connection NoMachine software,[13] a software that offers solutions for remote access. The access to the data is provided completely free of charge. External researchers gain access to a computing environment that does not allow users to transfer files to and from the server. They have access to a restricted area, where standard software such as *Stata*, *R*, *Julia*, and *Python* are available. The datasets for the project are placed in a read-only folder. Since there is no connection to the internet, the installation of specific packages must be requested through the BPLIM staff. The same holds true for the desired outputs. All outputs are subject to standard output control checks for confidentiality, and they are delivered to the researcher only thereafter. The duration of this process can vary, based on the complexity of the analysis and on the workload of BPLIM, with a potential time frame of up to a fortnight. However, it is typically completed within a span of 2 to 3 business days.

External researchers may also request to add external files to their accounts. The origin of these files must be clearly explained in a specific form and a justification for their use must also be provided. The researcher should also make sure these files can be legitimately used for that purpose. BPLIM staff will assist with the merging of the external datasets with BPLIM datasets. External datasets typically contain aggregated data, but it may be possible to add external datasets with finer granularity. In this latter case, additional measures may have to be undertaken to ensure that the confidentiality of the data is preserved. These situations will be analysed in a case-by-case fashion and will be discussed with BPLIM staff.

[13] www.nomachine.com

External researchers may be granted "surrogate access" to datasets that are normally not accessible to them, under specific circumstances. In this case, the researchers prepare scripts based on pseudo-data, which are then executed by BPLIM staff on the original data. No access to the external server is required in this case. This access mode is only possible for projects in which the topic is aligned with Banco de Portugal's research agenda. Thus, in order to secure surrogate access, external researchers must submit a comprehensive project proposal to BPLIM. The proposal will be evaluated according to the relevance of the topic for the research agenda of Banco de Portugal. Only projects deemed relevant will be granted surrogate access. If BPLIM decides to support the project, then the external researcher will be assigned a data expert from the BPLIM team who will interact with them and will help prepare and run the scripts on the original data. However, the researcher remains solely accountable for coding, with BPLIM refraining from validating or certifying scripts. Close collaboration with BPLIM staff is advised in order to ensure desired outcomes, and external researchers are encouraged to discuss findings with BPLIM staff. Initiating the project with a short-stay visit to the BPLIM is highly recommended and allows for discussions with staff and for the comprehension of data complexities. Other visits, during the course of the project, are also encouraged.

Outputs shared with external researchers adhere to standard disclosure restrictions, with access conditions for external researchers remaining unchanged under any circumstances. Furthermore, researchers are strongly encouraged to share their research with Banco de Portugal staff and to receive feedback throughout their project.

BPLIM retains only copies of the syntax files and datasets not originally prepared by BPLIM after the conclusion of a project. Datasets prepared by BPLIM remain replicable at any given time. It is the researcher's responsibility to ensure that all necessary materials for the accurate replication of results are archived within the project. Archives are maintained for ten years following the project's closure.

All BPLIM datasets should be cited following the guidelines outlined in each dataset manual. When available, the Digital Object Identifier (DOI) should always be referenced, as per the Service Level Agreement with GESIS (Fig. 23.2).

3.3 Most Relevant Datasets Available

All datasets are available in Stata format and are stored efficiently to minimise file size. Labels are applied to all variables, with value labels provided for categorical variables. Labels can be displayed in both Portuguese and English, whenever possible. Datasets contain unique unit identifiers, common across datasets within each project. Each dataset is accompanied by a manual and a set of auxiliary files that succinctly describe the data. Additionally, BPLIM datasets may include companion script files that calculate additional variables or harmonised variables to ensure comparability

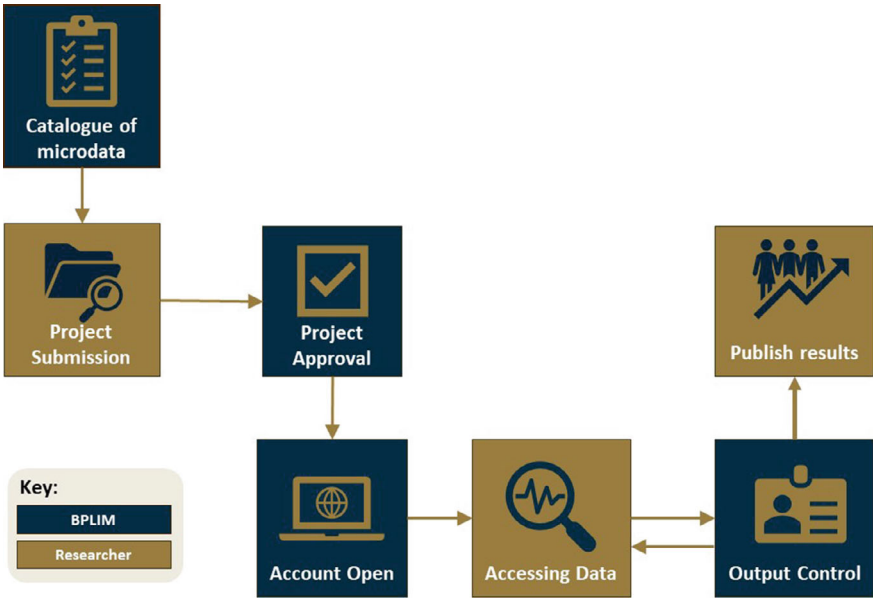


Fig. 23.2 Workflow of a Research Project

over time and across datasets. Regular updates are made to datasets based on a data extraction ("data freeze") at a specific point in time, with a versioning system being applied in order to reflect any changes to the dataset.

The list of the latest version of each dataset, along with all updated information including manuals and supporting documents, is always available on the BPLIM website. More detailed information for each dataset (codebook and descriptive statistics) can only be obtained once the researcher is given access to the server. Figure 23.3 shows the datasets that are currently available to external researchers.

The **Central Balance Sheet Database (CB)** provides economic and financial information on non-financial corporations operating in Portugal. This dataset is primarily sourced from information reported through the "*Informação Empresarial Simplificada*" (Simplified Business Information) and contains annual data from 2006 onwards.

The **Fast and Exceptional Enterprise Survey (COVID-19-IREE)** was launched by Statistics Portugal (INE) and Banco de Portugal with the aim of identifying the main effects of the COVID-19 pandemic on key aspects of enterprise activity, such as firm turnover, workforce, prices, credit conditions, and the use of government support measures.

The **Incentives Systems Database (SI)** contains information about projects submitted to incentive systems financed by the European Regional Development Fund (ERDF) under the QREN framework (2007-2013) and the European Regional

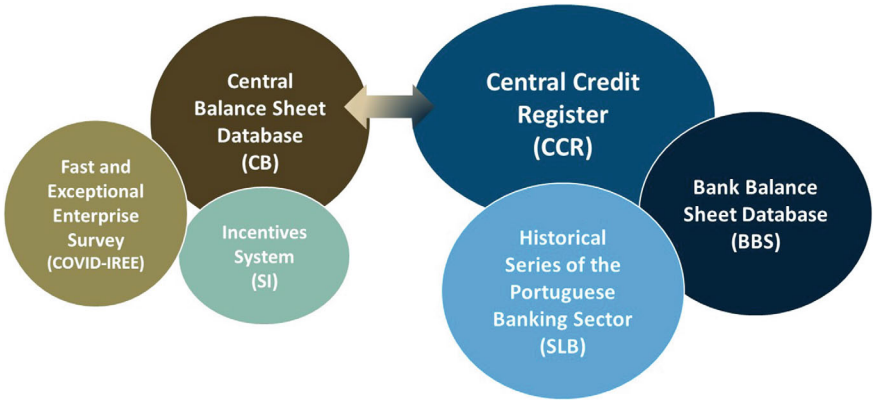


Fig. 23.3 Datasets available for External Researchers

Development Fund and European Social Fund (ESF) under the PT2020 framework (2014-2020). This dataset includes information about both approved and non-approved applications.

The **Central Credit Register (CCR)** database reports credit supply by all credit-granting institutions in Portugal, spanning from January 1999 to August 2018.

The **Bank Balance Sheet Database (BBS)** provides detailed information about the assets and liabilities of all monetary financial institutions (MFIs) operating in Portugal. This dataset contains monthly data from 1997 onwards.

The **Historical Series of the Portuguese Banking Sector Database (SLB)** provides, on a consolidated basis, a wide range of series on banks' financial statements (including balance sheets, income statements, and solvency), loans to customers, interest rates, human resources, branch networks, and payment systems. This dataset is updated annually and includes data dating back to 1990.

3.4 Data Confidentiality

Given that some datasets made available by BPLIM contain highly sensitive information, it became necessary to implement a data access solution that ensures the confidentiality of the data. All databases made available by BPLIM are at least pseudonymised, with the direct identifiers being stripped and a unique identifier replaced by pseudonyms: identifiers obtained artificially, which can act as masked representations of the original data. This process is implemented for both internal (affiliated with Banco de Portugal) and external researchers. To reinforce the anonymisation process, BPLIM implements a rotation in pseudonymisation to ensure that unique unit identifiers are common across datasets within each project, thereby

resulting in different identifiers for the same unit across projects. Thus, a researcher cannot link databases from two different projects. Additionally, the datasets are subject to a confidentiality classification as follows: **low**, **medium**, or **high**. If the variables in a dataset are classified as **low,** then the data is only pseudonymised. However, if a dataset is classified as **medium**, then the dataset will contain variables with "modified" values instead of the original ones. Datasets classified with a **high** level of confidentiality contain variables with "modified" values and are restricted to a subsample of the observations. BPLIM uses 3 strategies to create "modified" datasets:

- **Perturbation**—original sensitive values are replaced by different values, but some correlation with the original values is preserved;
- **Shuffling**—the original values of the sensitive variables are being kept but they are shuffled randomly and independently from the rest of the dataset to break the link between the observations. Noise can also be added in some cases (perturbed and shuffled);
- **Randomisation**[14] (pseudo-data)—randomly generated variables without any correlation with the original ones but respecting their metadata. The only purpose is to illustrate the structure of the data and to help build the scripts when the researcher cannot have access to the data.

BPLIM has developed tools to create the modified data. Tools for the creation of pseudo-data are publicly available at BPLIM GitHub.

The researcher prepares their scripts on modified data in order to produce the outputs required for the project. Since these outputs are obtained from the modified data, they do not contain valid results for research purposes. However, once researchers complete this task, they can ask BPLIM staff to rerun their scripts, this time using the original confidential data. These outputs are then subject to standard output control checks for confidential data and are then delivered to the researcher.[15] More specifically, the responsibility of running the scripts on the original data falls upon the internal researcher in projects co-authored with internal researchers, since he/she has access to the pseudonymised data. However, that responsibility falls upon BPLIM staff in projects placed on the external server. Since this is an exercise concerning reproducibility, it is important to guarantee that researchers follow the best practices in reproducibility.[16] One relevant aspect of this process is the use of containers that freeze the computational environment and that are specific to each project, which is strongly encouraged, particularly for open-source software.

[14] The term random arises from the fact that fictitious values are randomly generated, preserving some of the main statistical measures.

[15] See the *Harvard Data Science Review* article "Reproducibility With Confidential Data: The Experience of BPLIM" (https://hdsr.mitpress.mit.edu/pub/rl6maajr/release/1)

[16] See the following *Medium* article for an application of BPLIM's tool for creation of pseudo-data: https://medium.com/the-stata-gallery/the-bplim-workflow-for-anonymizing-confidential-data-for-research-6ea4c0d01396

BPLIM has developed a Python tool, the "Replication App," to ease the process of rerunning the code on the original data. The Replication App was designed for researchers who use the BPLIM's (modified) confidential datasets and runs on the external server. The code for the Replication App is available on BPLIM's GitHub, including versions for other operating systems. The researcher must first validate his/her code by successfully submitting the scripts through the Replication App before requesting a replication from BPLIM using the original data. The process involves selecting the main script, as well as all the required dependencies, created by the researcher. The folder structure used by the researcher is replicated and the BPLIM datasets must be read from the (read-only) data folder. All intermediary output files must be created during the replication. The researcher then uses the application to run the code. The code must run from top to bottom and produce no errors. If the run is successful, then the implementation on the original data requires only that BPLIM staff change the relative paths to the data folder and rerun the code. A side benefit of this process is that it automatically produces a replication package for the researcher. Stored in the folder are all replication scripts, the output files, as well as two additional files: one fully characterising the software environment, and another JSON file containing a listing of all scripts used in the replication. If the definition file that was used to produce the container is added (or a listing of all packages and their versions), then the complete replication package (except the data) is ready. This package is provided to the researcher, and BPLIM gives them the permission to distribute it for public access if they wish.

4 Conclusions

Access to administrative microdata in Portugal occurs through two main statistical authorities, namely Statistics Portugal and Banco de Portugal. At the time of writing, Statistics Portugal is successfully integrating administrative data sources to improve official statistics. Key benefits include enhanced analysis, efficiency, and innovation. Challenges include delays and data quality concerns. The organisation is addressing these challenges through partnerships, streamlined access, and data security measures. Moreover, one of the key objectives is the provision of research access to anonymised microdata.

In this context, Statistics Portugal envisages the following:

- Increasing the number of safe centres by locating remote safe centres in universities and in other research institutions or departments;
- Increase the automatisation of the output check;
- Implement sampling and anonymisation procedures to decrease the identification probability and to provide microdata versions accessible outside the safe environment, by transforming the dataset into a scientific use file;
- Increase the number of those who are given access to data for research purposes;

- Implement a coding procedure to allow researchers to integrate the different administrative data sources when the observation unit is the individual;
- Support researchers' or other players'—e.g., public administration institutions—initiatives and partnerships that promote data scouting and who are willing to serve as a data hub for research;
- The centralisation of all possible administrative microdata at Statistics Portugal should increase, given the research potential of data integration. The centralisation should also include datasets currently accessed for research from other statistical authorities (even if duplicates access) or microdata that do not necessarily feed official statistics.

Meanwhile, a complimentary highly developed data access point is the Banco de Portugal, through its Microdata Research Laboratory (BPLIM). Its data is accessible to researchers worldwide, following clear and well-documented data access procedures. Since BPLIM provides highly confidential data, it has become an imperative to establish a data access mechanism that guarantees the privacy and security of the information.

Banco de Portugal is constantly working on improving access to its data for research purposes. It has plans to expand the list of datasets available to external researchers, including both national and international datasets. Regarding international datasets, there are several recent initiatives involving collaboration with other central banks, international institutions, and national statistical institutes to share information and experiences on effectively handling granular data and making it accessible to external users.

Within this realm, a specific initiative in which Banco de Portugal is actively involved is the **I**nternational **N**etwork for **E**xchanging E**x**perience on Statistical Handling of Granular **Da**ta (INEXDA).[17] One of this network's aims is to look into harmonising the procedures that govern access to granular data across countries. Banco de Portugal has been one of the founding members of the network, actively engaging in the exchange of ideas across different institutions across Europe and the rest of the world. While the harmonisation of access across countries is still in its infancy, initiatives like INEXDA are an important step towards achieving this goal.

Francisco Lima has been President of Statistics Portugal (INE) in 2018-2024. Professor at the Instituto Superior Técnico, University of Lisbon, since 2003. Doctorate in Economics, Universidade Nova de Lisboa (2001). His research focuses on applied microeconometrics, labour economics, entrepreneurship, innovation and policy evaluation. He has coordinated and participated in several national and European research projects and has published in international journals such as the Journal of Political Economy, Economic Journal, Research Policy, European Economic Review, Small Business Economics and Labour Economics.

[17] www.inexda.org

José Pinto Martins is the Director of the Dissemination Unit at Statistics Portugal since 2007, with over 30 years of experience in the institution. He holds qualifications in Electrical Engineering, Information Management Science, and Statistics. He has led innovative projects in statistical dissemination, coordinated major conferences such as Q2024 and ISI 2007, and contributed to advancing statistical literacy. A frequent collaborator with Eurostat and other organizations, he focuses on improving statistical communication, quality standards, statistical literacy and accessibility within the European Statistical System.

Katja Neugebauer is a Senior Economist at the Financial Stability Department of Banco de Portugal. Her prior professional experience includes work as a Researcher at the Systemic Risk Centre at The London School of Economics and Political Science and as a Researcher at the Finance and Economics Unit at the Joint Research Centre of the European Commission. She holds a Master's degree in Quantitative Economics and a Ph.D. in Economics from the University of Tübingen. Her current research interests focus on residential real estate, access to bank credit, and crypto-assets and Decentralised Finance (DeFi). Her work has been published in peer-reviewed journals such as the Journal of Banking and Finance, Journal of Financial Stability, Economic Policy, Journal of International Money and Finance, and Empirical Economics.

Joana F. Pimentel has been a Senior Adviser at the Economics and Research Department at Banco de Portugal since June 2015. Previously, at Banco de Portugal, she spent eight years at the Prudential Supervision Department, where she participated in the implementation of Basel III in Europe (focusing on liquidity measures) and in the European legislation process on securitization, both in cooperation with the European Banking Authority and the European Central Bank. Before joining Banco de Portugal, she was a Risk Consultant at Deloitte & Touche in Boston and an Auditor at Deloitte & Touche in Porto (formerly Andersen). She received her Master's degree in Finance from the Carroll Graduate School of Management – Boston College (2005) and her undergraduate degree in Management from Faculdade de Economia do Porto (FEP, 2000).

Sofia Rodrigues is Head of the Administrative and Business Data Unit, in Data Collection and Management Department at Statistics Portugal. She has extensive experience in several statistical fields, including National Accounts; Planning; Business Registers and Business Statistics. Since 2019, she has been responsible for managing the team focused on collecting and analysing data from companies and institutions through online surveys, and administrative and other sources. Her work supports the production of official statistics and the provision of reliable and accessible data for scientific research.

Paulo Saraiva is Director of Data Collection and Management at Statistics Portugal, brings extensive experience in integrating data from business and household surveys, censuses, and administrative sources. He has driven modernization initiatives in data collection, delivering high-quality and timely outputs to support economic and social policy. His efforts centre on harmonizing statistical practices, optimizing resource use, and fostering collaboration to improve data accuracy and usability across diverse areas.

Rita Sousa holds a Ph.D. in Statistics and Risk Management from the New University of Lisbon. Since 2015, she has been a Senior Data Analyst at the Microdata Research Laboratory (BPLIM), within the Economics and Research Department of Banco de Portugal. Previously, she worked for 15 years as a Methodologist in the Department of Methodology and Information Systems at Statistics Portugal (2000–2015). She was also an Invited Assistant Professor at the University of Porto for over a decade and has extensive experience as a trainer in statistical software, particularly R. Her research interests include sampling, analysis of large datasets, estimation of sensitive variables using auxiliary information, and statistical disclosure control.

João Zilhão is Member of the Board of Statistics Portugal since 2018 and former Director of the Planning, Control, and Quality Unit (2007–2018), holds a degree in Economics (ISEG Lisbon) and a Master in Statistics and Information Management (NOVA IMS Lisbon). Active in Quality Management in Statistics within the ESS since 1999, she has led key EU quality initiatives, including Peer Reviews (2021–2024), and served as Chair of the EU Council Working Group on Statistics during Portugal's 2021 EU Presidency. She is an experienced trainer in quality in statistics and a contributor to international cooperation and conferences in the theme.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 24
Pathways to Accessing Administrative Microdata in Romania



Andreea Mitrut, Oana Borcan, and Alina Botezat

Abstract Access to microdata in Romania has improved over the course of the past two decades, largely due to investments in infrastructure made by the National Institute of Statistics (NIS) and on the basis of the increased availability of individual level data from certain public authorities. However, linked individual-level administrative datasets are still unavailable, despite the aforementioned progress. Only a few sectors, such as education and public health, maintain comprehensive and regularly updated datasets that are accessible. This chapter outlines the main data sources, highlights key challenges in accessing them and offers practical recommendations for researchers and institutions seeking to secure or to improve data access.

1 Introduction

Data collection and access in Romania have improved substantially over the past two decades, even though full linkage across data and accessibility of all individual and firm-level data have yet to be achieved.

This chapter's structure is as follows: *Section 2* provides a brief overview of microdata availability via the NIS in Romania and covers data types, structure, access, and key limitations, particularly for non-Romanian researchers. It concludes with a short description of the application process. *Section 3* outlines additional data sources that may be of interest to researchers. The information presented is based on the authors' experience and may not capture all existing datasets or access pathways. *Section 4* concludes.

A. Mitrut (✉)
University of Gothenburg, Gothenburg, Sweden
e-mail: andreea.mitrut@economics.gu.se

O. Borcan
University of East Anglia, Norwich, UK
e-mail: O.Borcan@uea.ac.uk

A. Botezat
Gheorghe Zane Institute for Economic and Social Research, Romanian Academy, Iaşi, Romania
e-mail: botezat.alina@yahoo.com

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_24

2 Microdata via the National Institute of Statistics in Romania

The National Institute of Statistics (NIS) leads efforts in systematic data collection across a wide range of economic and social indicators. The institute currently provides access to three types of data: microdata, aggregated data, and metadata. In this section we will focus on the microdata.

Microdata sets from NIS are available across various fields, including: **Population and Demographics** (registers on mortality, natality, nuptiality, and divorce rates); **Living Conditions** (household budget surveys, quality of life assessments, and access to information and communication technology (ICT) in households); **Labour Force** (employment-related surveys such as the Labour Force Survey, the Youth Access to the Labour Market Survey, and the Earnings Structure Survey); **Enterprise Statistics** (Structural Business Survey, Industry Business Outlook Survey, and Statistical Business Register); **International Trade** (Foreign Trade in Goods Survey); **Education** (limited microdata; only the Adult Education Survey is available for one year in Excel format); **Health** (only Health Status Surveys and Health Activity Surveys); **Research and Innovation** (Innovation Surveys); **Prices** (consumer prices and industrial production prices). Additionally, the NIS may also provide access to individual-level data from the national censuses conducted after 1992.

These databases are independent and most of them lack unique identifiers; moreover, to our knowledge, data merging services are not provided. Additionally, questionnaires and key documentation are only available in Romanian. The NIS grants access to statistical microdata for scientific research purposes exclusively. Access is restricted to individuals from accredited research institutions (e.g., universities, research institutes, public administration research departments, central banks, and international organisations). While foreign institutions can request access, their use is limited by the fact that contracts and data are typically in Romanian. The application details are outlined on the NIS's website and imply that the applicants must first submit an *official request* from their accredited educational institution or research institute.[1] The request must include the typical information other countries have when requesting similar data: the project description, including methodology and variables to be used, together with a brief overview of the expected results. Additionally, the request must be accompanied by the annexes specified in the *Application Guide of Regulation (EU) 557/2013 on European Statistics*, regarding access to confidential data for scientific purposes. These annexes are available on the NIS's webpage and can be submitted easily by mail or email. Access to these databases is free of charge, and the decision to grant access is based solely on the submission of the necessary documents, rather than on the scientific merit of the project itself.

[1] https://insse.ro/cms/en/content/access-microdata.

3 Other Relevant Sources of Microdata

Beyond the data available via NIS, there are also several other databases providing high-quality administrative, and generally accessible, microdata. Please note that this is not an exhaustive list, but rather a reflection of our experience of working with Romanian microdata.

(i) **Education** is one of the most accessible and frequently explored fields for data in Romania. Until 2022, detailed population-level data on high school admissions (grade 8) and national high school leaving exams (grade 12) was publicly available online (see http://static.bacalaureat.edu.ro/ and http://static.admitere.edu.ro (both in Romanian)). Alternatively, the Ministry of Education might give access to some disaggregated data, especially about participants in government programmes (see below), but there is no clear application process.

On the government's open data portal,[2] the Ministry of Education has been publishing a series of microdata databases since 2014. These include anonymised results from the National Evaluation—a standardised national high school admission exam—and the Baccalaureate—a standardised national high school exit exam—are available in CSV and XLSX formats. Additionally, the portal has provided data on the number of students enrolled in each school in Romania, available in XLSX format, since 2015.

Users can request datasets that are not available on the Portal by filling out an online form at https://data.gov.ro/datarequest. Public institutions may grant access to their documents in existing formats and in any available language. Whenever possible, data is provided in open, machine-readable formats along with metadata that describes the content.

In 2012, pseudo-identified pupil-level data was obtained through direct requests made to the Ministry of Education and were used in Borcan et al. (2017) in a paper that studied the impact of the introduction of CCTV cameras in national school-leaving exams on pupils' outcomes. The data obtained enabled the researchers to match exam outcomes across the high school admission and the Baccalaureate exams, as well as the pupil-level data on a means-tested cash transfer programme ("Bani de Liceu"), serving as a proxy for low-income students. This enabled the researchers to examine the policy's distributional consequences. Further relevant examples using education data from Romania include Borcan et al. (2014), Malamud et al. (2023), and Malamud et al. (2025).

(ii) **Health**: The National Institute for Health Services Management (www.insp.gov.ro) provides access to in-patient data from 2008 onward. It maintains a comprehensive national database that is available to researchers in a semi-aggregated format. The institute offers a smooth application process and excellent communication support. Data access is available at the hospital at a per-month level and includes standard patient characteristics, ICD-coded diagnoses, and procedures. While detailed documentation is limited, the staff has

[2] https://data.gov.ro/organization.

been very responsive and helpful via email. There is a fee for accessing the data, one that is in line with costs charged by similar institutions providing aggregated health data. Andreea Mitrut is currently involved in several research projects using this data.

(iii) **Civil servants and public officials' asset and wealth declarations**: This data is available at the individual level for all current employees in the public sector in Romania who serve in an executive post (or those who were employed in an executive post in the past five years). This includes top-level bureaucrats, local councilors, school and hospital managers, party executives, members of the legislative and government bodies. The data is searchable by the official's name, position, institution, county, municipality or date of publication of their declarations. Yearly reporting of assets and wealth is mandatory for all holders of executive roles; therefore, researchers could use this data source to compile longitudinal datasets. The data is available at https://declaratii.integritate.eu, the search engine is in Romanian, but is quite self-explanatory. The declarations are non-searchable pdf files (often handwritten and scanned copies of the declarations), and therefore research projects requiring this data could incur larger costs relating to digitisation. OCR software or research assistance would be essential to mass-digitise the data. One example of using this data source in political science research is Klašnja (2015), who digitised around 4000 mandatory declarations of assets, relating to mayoral candidates in Romania, in order to measure corruption through disproportionate increases in wealth. The data can also be used to identify the political affiliation of public officials in executive positions, with applications, such as research on the political connections of different institutions and conflicts of interest.

In addition to these sources of disaggregated administrative data, there is also Electoral **data**, recorded at various levels of aggregation**.** It is publicly available from the Central Electoral Bureau (BEC) via the Romanian Electoral Authority (AEP) (it is also available in English). However, upon selecting the round of elections of interest, the visitor is redirected to BEC for data for that election (e.g., https://parlamentare2020.bec.ro/rezultate). These webpages are not available in English, making access more difficult for foreign researchers. The data, such as turnout by polling station or aggregated at municipality level or number of valid votes for different candidates or parties was available in an editable format (e.g., csv) until 2016. However, from 2020 onwards, the results are presented as aggregate statistics at the national level, or were displayed in uneditable formats, thereby adding further complexity to the data collection process.

4 Concluding Remarks

To conclude, researchers face two key challenges in accessing microdata. First, data requests generally need to be submitted in Romanian, in order to increase the likelihood of a favourable response, and survey questionnaires and variables are also only available in Romanian. Second, Romanian authorities are often reluctant to share data, due to concerns over GDPR compliance, even when certain nominal data—such as the identities of individuals in managerial or executive positions within state administration, including schools—is legally considered to be public interest data.

In December 2024, Romania adopted its first national open data strategy to align with EU and OECD recommendations. Formerly the only EU country without such a strategy, Romania faced major gaps in open data management, transparency, and infrastructure. The new strategy aims to improve accessibility, legal frameworks, staff training, and data usability.

References

Borcan, O., Lindahl, M., & Mitrut, A. (2014). The impact of unexpected wage cut on corruption: Evidence from a 'xeroxed' exam. *Journal of Public Economics, 120*, 32–47.

Borcan, O., Lindahl, M., & Mitrut, A. (2017). Fighting corruption in education: What works and who benefits? *American Economic Journal: Economic Policy, 9*(1), 180–209.

Klašnja, M. (2015). Corruption and the incumbency disadvantage: Theory and evidence. *The Journal of Politics, 77*(4), 928–942.

Malamud, O., Mitrut, A., & Pop-Eleches, C. (2023). The effect of education on mortality and health: Evidence from a schooling expansion in Romania. *Journal of Human Resources, 58*(2), 561–592.

Malamud, O., Mitrut, A., Pop-Eleches, C., & Urquiola, M. (2025). Self-, peer-, and teacher perceptions under school tracking. *The Review of Economics and Statistics*, 1–41. https://doi.org/10.1162/rest_a_01555

Andreea Mitrut is a Professor of Economics at the University of Gothenburg. She is an applied microeconomist with research interests at the intersection of labor economics, health economics, economics of education, and political economy. She is interested in topics related to education, early child development, fertility and family planning, corruption, discrimination and social exclusion. She has extensive experience working with administrative microdata from Romania. Andreea has published in journals such as the Review of Economics and Statistics, American Economic Journal: Economic Policy, the Economic Journal, Journal of Human Resources, Journal of Public Economics and Journal of Health Economics.

Oana Borcan is an Associate Professor of Economics at the University of East Anglia. She is an applied economist working in development economics and impact evaluation, focusing on corruption, institutions and human capital. She has worked extensively with Romanian administrative data on education. Oana is an Abdul Latif Jameel Poverty Action Lab invited researcher and has published in journals such as American Economic Journal: Economic Policy, Journal of Economic Growth and Journal of Public Economics.

Alina Botezat is a Senior researcher at the Gheorghe Zane Institute for Economic and Social Research of the Romanian Academy in Iasi. Her research interests span the fields of economics of education, migration, policy impact evaluation, and the broader domains of wellbeing and health. She is based in Romania and has extensive experience working with Romanian data. Alina has published in journals such as the Journal of Economic Psychology, World Development, Population, Space and Place, and Globalization and Health.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 25
Slovak Administrative Data Increasing the Value for Money in Public Spending



Miroslav Štefánik

Abstract Slovakia is a country with a solid tradition in administrative data collection. Procedures for accessing individual-level statistical data follow the principles of the European Statistical System, but do not cover access to administrative data; these are instead collected and maintained by responsible public bodies. The capacities of these public bodies often limit the facilitation of access to administrative data. This situation has been changing since 2016, when particular ministries started to build their analytical units under the government-wide initiative of "Value For Money" in public service. Newly built units started producing policy-relevant analytics intensively, often based on administrative data. We walk through this development in particular policy domains, explaining their data context and picking from among examples when administrative data-based evidence was translated into either a policy decision or the adjustment of an existing policy tool. We select examples of the different forms of translating analytics to Slovak policy-making, documenting: i. the introduction of a new policy measure; ii. contribution to a more efficient set-up of existing policy measures; or iii. steering the public debate towards relevant issues.

1 Introduction

This chapter outlines the administrative data utilisation in Slovak decision-making. The extent of this chapter does not allow for an exhaustive list of instances in which administrative data (AD) based evidence was translated into a policy decision. Instead, we aim to select some of the analyses with a transparent link between their conclusions and a policy decision. Moreover, in outlining the practices selected, we try to document various forms of translating evidence into policy-making. We reference examples where AD-based evidence led to the introduction of a new, previously non-existent policy, to the adjustment of an existing policy measure, or when it contributed to framing the public debate and public policy's top priorities.

M. Štefánik (✉)
Institute of Economic Research of the Slovak Academy of Sciences, Bratislava, Slovakia
e-mail: miroslav.stefanik@savba.sk

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_25

The first section describes the relevant context of Slovak administrative data collection, which is crucial to understanding the character and intensity of analytical production. The second section outlines selected examples of AD-based analyses and is grouped by policy domains. The third section concludes and sets the examples presented in a broader context.

2 Country Context

Slovakia's accession to the European Union in 2004 concluded a tumultuous century in the nation's quest for recognition. In 1918, the Slovaks, one of the many nations of the ethnically heterogeneous Austro-Hungarian Empire, merged with the Czechs to form Czechoslovakia. After the Second World War, Czechoslovakia fell into the bloc of socialist countries, only to reassume democratic governance in 1989. Slovakia gained independence shortly thereafter, after a peaceful split with Czechoslovakia in 1993. Its dynamic, modern history determines the shape of the country's governance.

2.1 Tradition of Data Collection

A long tradition of collecting administrative data is rooted in the heritage of the Austro-Hungarian administration, thanks to which the Slovak population has been covered by censuses since 1787.[1] The first Czechoslovak Republic, emerging in 1918, proved to be a progressive and economically strong state, with public administration employees enjoying a high social status. Between 1948 and 1989, the Czechoslovak-planned economy demanded more intensive data collection, not only on demography but also on economic production; the domains of agricultural and industrial production were densely covered, but so too were crime and security; solid administrative data were also collected on public services such as health or education. Prior to 1989, unemployment and poverty did not officially exist; administrative data collection covering these domains had to be built anew during the 1990s. The distribution of unemployment or social benefits demanded operating registers, which had to be introduced swiftly, often at the expense of data quality.

Additionally, the income gap between public and private sector employees began to rise and persisted at a considerable level during the 1990s and has lasted until today (Ministry of Finance of the Slovak Republic, 2018). The long-term income gap has negatively impacted the public administration's competency, often reducing its

[1] https://slovak.statistics.sk/wps/portal/ext/aboutus/history/!ut/p/z1/jZDLDoIwEEW_hS_ohYKty1ZDqSlgIS3YjWFlSHwtjN-vUVcmIrOb5Nw5k0sC6Uk4D_fxMNzGy3k4PvddWOy92XIpYwFkRQ6bNap0safWJ6R7ASslipQZgBuVQYvCNUtLKQQlYU4eP0ZgXl5Ya1vjPZRPcmgaK1TOAeVM_wQQpt_rSHghlun3BV6tJTQzzHItaS2TDzBV0T_Jl6IuFJ6KTZs2nCWIM3I9Oed6jPoQRQ8IXQvF/dz/d5/L2dJQSEvUUt3QS80TmxFL1o2X1E3SThCQjFBME9IRzAwSTdKUzRSODcyMFA3/.

analytical duties to formal, ex-ante impact assessments of new legislative proposals. An initiative, entitled "Value for Money," has aimed to compensate for this trend since 2016.

2.2 *Value for Money Initiative*

The Slovak Ministry of Finance harboured the most recognisable analytical unit prior to 2016, much like in many other European countries. The concept[2] of value for money in public services spread upwards to be adopted in the Programmes of the Slovak governments between 2016 and 2023, based on an initiative originating in this precise intellectual environment. Adoption at the highest levels of decision-making opened options for additional funding, which were used to build:

(i) a new Value for Money Unit (VMU) of the Ministry of Finance of the Slovak Republic (SR, since 2016);
(ii) analytical units (AU)/Institutes at the Slovak Ministries and the Governmental Office.

In its initial stages, and up until 2020, the initiative was dominantly covered by EU funding under the Operational Programme Effective Public Service.[3]

In the first step, the VMU also cooperated with respective AUs, where applicable, to prepare spending reviews according to policy areas. The spending reviews followed a prescribed methodology and included a list of proposed policy measures aimed at increasing the value of money spent in each of the policy domains being reviewed. Later, the focus moved to ad-hoc thematic studies addressing topics relevant from the perspective of the Slovak economy, by applying benchmarking comparisons of the performance of Slovak public services to other countries. Finally, since 2020, the VMU has been assigned to assess all public investments of more than one million euro.[4] Establishing the VMU and AUs presented a historical push to the administration, maintenance and utilisation of administrative data. Through the Ministry of Finance, the VMU gained and maintained access to all administrative data collected by the Slovak public sector, physically storing and updating data from the most relevant sources.

In addition to data access, VMU and AUs have also benefited from access to training and decent salaries that were comparable to the private sector average. Since 2020, VMU with AUs have been standardising their quality assurance process, formalising the commenting and revision process of their analytical reports under a joint system coordinated by the VMU. The most essential outputs are published

[2] The assessment of public expenditures to gain the highest value out of public spending. The original concept mentioned three methodologies: Spending review, Benchmarking (international comparison), and Cost-Benefit Analysis (Source: https://www.trend.sk/spravy/najlepsi-moznych-svetov-hodnota-peniaze-slovenskej-verejnej-politike).

[3] https://www.mfsr.sk/sk/projekty-mfsr/projekty-op-evs/.

[4] https://e.dennikn.sk/1806229/otazka-za-milion-eur-co-je-hodnota-za-peniaze-a-co-uz-nie/.

centrally in a joint repository (https://analyzy.gov.sk/). The Value for Money initiative shifted the ability of the Slovak decision-making sector to a qualitatively higher level.

2.3 The Legal and Organisational Framework of the Data Access

Administrative data sources have traditionally been a part of the state statistics in Czechoslovakia and continued after Slovakia acquired independence in 1993 and joined the European Union with the European Statistical System.[5] Under the Slovak System of State Statistics,[6] the Statistical Office of the Slovak Republic (SO SR) annually issues a Decree defining the data requests made to particular institutions that are obliged to report (aggregate or individual level) information based on data collected under their administration. These data flows fill a comprehensive list of policy-relevant statistical indicators. A typical example is the unemployment rate, which is based on the register of unemployed persons. The internationally less common examples include sector-specific data on disease incidence, air and water quality, or childcare provision.

The most comprehensive administrative databases (registers of private companies and physical persons[7]) also serve as the initial base for sampling statistical surveys. Additionally, the utilisation of administrative databases in combination with CENSUS-related data collection is of particular interest. Administrative data is partially used to collect CENSUS data in Slovakia, with the registers being used as the main base for non-response imputations. Despite the declared intention for a wider utilisation of administrative data in CENSUS 2021, a shift to a CENSUS entirely based on administrative data sources has not yet been introduced (Vaňo, 2014).

Implementing the EU-wide GDPR (Regulation (EU) 2016/679) in the Slovak legislation reshuffled the data access options for all of the actors involved in policy evaluation, academia, and analytical units in the public sector. Institutions collecting and administering the data are only eligible to provide the data for purposes defined in the existing legislation. While particular Ministries, pioneered by the Ministry of Finance, managed to update the necessary legislation defining the character and purpose of required data exchange, academic bodies lacked the mandatory legislative coverage. In practice, academia either relies on the limited (mostly aggregated-level) data exchange offered through the SOSR or has to shelter access to focused administrative data under bilateral, project-based cooperations with particular data-collecting institutions or AUs.

[5] https://ec.europa.eu/eurostat/web/european-statistical-system.

[6] This was legally defined by the Act on State Statistics (No. 540/2001, Coll.).

[7] While the Register of Private Companies is directly maintained by the SO SR, the Register of Physical Persons is maintained by the Ministry of Interior of the Slovak Republic.

Ultimately, access to AD twists with the level of aggregation. Aggregated information from AD can be requested from all public bodies by all physical persons (regardless of their citizenship or nationality) based on the Act on Free Access to Information (No. 211/2000 Coll.). Public bodies can reject an information request submitted under this Act only when the information requested is either (legally) classified or personal (violating the GDPR principles). Since a dominant share of the AD collected by public bodies includes personal data, access to individual-level databases crucially rests on the ability and willingness of the particular public institution to anonymise and prepare the requested data export. In practice, the presence of personal information in the original, individual-level database often serves as an excuse from the obligation to provide data under the Act on Free Access to Information. In practice, data requests submitted under this Act, therefore, involve at least some level of aggregation from the original AD.

In cases for which individual-level AD are necessary, these might be provided for research purposes[8] exclusively by the SOSR, based on the Act on State Statistics (No. 540/2001, Coll.). While the most relevant AD are a part of the State Statistics, and SOSR maintains the necessary data flows from responsible AD collecting institutions, it does not facilitate access thereto for research purposes. SOSR facilitates recognised research entities' access to individual-level data through a standardised procedure, but only from statistical surveys collected by SOSR directly. The SOSR does not offer a standardised procedure in the case of AD requests submitted under the Act on State Statistics; rather, it instructs the requesting research institution to submit requests to the AD-collecting institution. Additionally, while data from statistical surveys collected by SOSR are well documented by metadata, no metadata beyond simple database descriptions are available for the AD sources. The AD-collecting institutions do not usually feel obliged by the Act on State Statistics, since it operates under its specific legislation.[9] Thus, they usually claim an insufficient legal basis to provide sensitive (personal) data or a lack of capacity to generate an anonymised data export in response to such requests. This creates a bottleneck situation in which public bodies collecting AD claim not to have the capacities and competency for anonymisation, and the SOSR, while actually having the competency, points to the absence of a legal basis to provide data not collected by SOSR. The few examples of access to individual-level AD granted to research institutions for research purposes were based on ad-hoc cases when researchers convinced the AD managing institution to employ their internal capacities in generating an anonymised data export for their research.

The exchange of AD, including personal data, is allowed and frequently executed at the level of ministries. The Ministry of Finance, with VMU and its underlying AU, benefits from a central position in this data exchange network, thanks to a legal basis allowing the exchange of individual-level AD, possibly even containing

[8] Concerning research institutions recognised by SOSR or Eurostat: https://ec.europa.eu/eurostat/documents/203647/771732/Recognised-research-entities.pdf.

[9] e.g. the Social Insurance Agency operates based on the Act on Social Insurance No. 461/2003 Coll., etc.

personal and sensitive information, by adding one point (§14, point 5) into the Act on Budgetary Rules of Public Administration (No. 523/2004, Coll.). Based on this point, all subjects drawing state budget funding are obliged to provide information (inc. personal) to the Ministry of Finance for the purpose of both creating and evaluating the state budget with its particular spending items. This point assures exclusive AD access to the employees of the Ministry of Finance, including its analytical teams under both the original AU and the more recently created VMU. Subsequently, other Ministries followed the example of the Ministry of Finance in their respective areas of operation. The possibility of working with personal data (including the universally used personal IDs) also enables linkages between various AD.

As a result, the production of ministerial analytical units has outshined academia in terms of quantity and policy relevance. However, the difference was partially driven by academia's (relatively less favourable) position in requesting individual-level AD for research purposes. While there are long-term operating legal and organisational schemes of data exchange between the AD collectors and academia for data aggregated from AD, as well as individual-level data from statistical surveys, the provision of individual-level AD critically rests on the limited willingness of the AD collector to prepare an anonymised data export. In reaction, academic institutions had to focus more on producing studies that were publishable in international scientific journals, and there were only a few examples of collaboration between the ministerial analytical units and academic institutions. These document the possibility of enhancing the undisputably positive impact of processing country-specific, individual-level AD by employing the underutilised capacities of academic researchers. One possible way forward involves the initiatives generating documented, AD-based evaluation datasets that are widely accessible for research purposes, such as the evaluation datasets produced by the German Institute for Employment Research.[10]

3 Utilisation of Administrative Data in Evidence-Based Policy-Making

In this section, we examine selected examples of translating the AD-based evidence into adjustments to existing policies or regulations. The examples are organised by policy domains, describing the most relevant and utilised administrative data sources and highlighting the best examples of translating an AD-based analysis to policy decisions. The list presented does not aim to provide a complete outline; instead, it seeks to provide examples of different types of AD-based evidence that translate into policy-making. Specific focus is paid to findings based on impact evaluations, but we do not restrict our attention to this family of studies.

[10] https://fdz.iab.de/en/our-data-products/archived-data/aed-iza-iab-administrative-evaluation-dataset/.

3.1 Labour Market Policy

The domain of labour market policy rests on two "backbone" administrative data sources: i. the register of unemployed persons complemented by their participation in active labour market policy (RUP) programmes; ii. the register of employed persons contributing to the Social Insurance (SI) system. No legal unemployment was formally possible in the planned economy prior to 1989. The register of unemployed persons, along with the system of unemployment insurance and active labour market policy (ALMP), had to be built entirely anew in 1990, following the experiences of other European countries.

Despite their initial development stage, RUP data covering 20 Slovak districts were processed in a few early studies identifying a positive impact of active labour market policy programmes (Lubyova & Van Ours, 1998 and 1999). Lubyova and Van Ours also analysed the impact of changes made to the unemployment insurance system on unemployment dynamics (1997). These studies present the first impact evaluation studies in the context of Slovak employment policies; this was followed by a longer silence caused by restricted access to the RUP database. During this period, the analytical unit of the Ministry of Finance used the Labour Force Survey (not the first-choice dataset) to generate evidence about the employment impact of Slovak ALMPs (Harvan, 2011). In 2014, another export from RUP and SI was generated, this time resulting in a chain of studies estimating the impact of ALMPs on the labour market.[11] With the exception of two studies (Harvan, 2011 and Hidas et al. 2016), a dominant share of ALMP impact evaluation studies originated from the academic sector. The situation started to change after 2018 when the Ministry of Labour, Social Affairs and Family (MoLSAaF) analytical unit was established under the Value for Money initiative. The analytical unit produced impact evaluations of five ALMP programmes (Petráš, 2018, 2019, 2020, 2021; and 2023) and employment protection measures implemented during the COVID-19 Pandemic (Bělín & Veselková, 2022; Vaľková & Petráš, 2022).

In Slovakia, like in other countries, labour market policy presents a policy area most densely documented by impact evaluation studies. Despite the long-term coverage, only a few instances of analytics in policy-making can be highlighted. The studies emerging from the academic sector are often not referenced by a particular policy decision; in such cases, we remain with only a timing sequence linking the decision to an analysis. For example, a study documenting the cost-effectiveness of existing spatial mobility support schemes (Štefánik & Karasová, 2016) was followed by an unprecedented increase in the budget allocated to these measures in 2018.[12]

The story of training-oriented ALMPs offers another interesting example. The very early impact evaluation study (Lubyova & Van Ours, 1999) estimated a positive impact of the re-training programme on the transition from unemployment.

[11] An overview of the studies is provided in Štefánik (2018); for more recent studies see e.g. Švábová et al. (2022) or Štefánik and Lafférs (2024).

[12] https://www.podnikajte.sk/financie-na-podnikanie/zvysenie-prispevkov-dochadzku-pracou-01-06-2018.

After almost two decades, impact evaluation studies have pointed to their non-significant or even negative impact on post-participation income or employment chances (Lubyová et al., 2014; Štefánik, 2014); they have also highlighted the undernourishment of training programmes in the portfolio of Slovak ALMPs, potentially driven by particular problems in the delivery of training (Hidas et al., 2016). After 2014, training delivery started to shift to a more flexible model with the arrival of client-driven choices of training specialisation and provider. Subsequent impact evaluation studies document that this change moved the training's impact from negative or non-significant to moderate, but positive and statistically significant figures (Petraš, 2018; Stefanik, 2021). The story of the evaluation of the ALMP-provided training offers an example when the impact evaluation results needed to be put into context and supported by a more in-depth analysis of the underlying mechanisms. The negative effects estimated for the training provided before 2014 could have led to a restriction or even abolishment of the only formal training programme for the unemployed in the ALMP portfolio. Instead, they stimulated a reform of training provision, thanks to international benchmarking (of the total resources flowing to this type of programmes) supported by an in-depth analysis.

The experience of the public works type of programme took the less favourable alley, with the programme being temporarily abolished based on negative results from impact evaluations (Bořík, 2015; Lubyová et al., 2014; Hidas et al., 2016; Petráš, 2020) despite the fact that some of the studies pointed at a high heterogeneity behind the negative or non-significant average treatment effects (Lubyová et al., 2014) or more favourable results of the programme, considering it is usually provided later in the unemployment spell (Štefánik & Lafférs, 2024). Despite a wide critique from the side of non-governmental organisations and municipalities, MoLSAaF decided to abolish the programme in 2024 without a suitable substitute programme targeting the long-term unemployed, those most remote from the labour market. The negative impact evaluation results were used to justify the easier solution of abolishing, instead of reforming, the programme. Later, in 2024 and in reaction to the critique of the municipalities and stakeholders in the field, MoLSAaF decided to renew the public works programme under a restricted form.

In Slovakia, as in other countries, the employment policy domain, especially ALMP, has a rich tradition of impact evaluation studies. Therefore, it brings various examples of translating the results of the impact evaluation studies into policy practice.

3.2 Educational Policy

The education system in Slovakia is comprehensively covered by the so-called Resort Information System (RIS), aggregating information from particular schools at all levels of education. The RIS collects individual-level information about the number of pupils and students, as well as teachers and support staff. In addition to RIS, there is data collection related to inspecting the educational outcomes of students at

various stages of their educational tracks, e.g. before leaving the elementary school, at the entrance and graduation from secondary schools. At the university level, data about educational outcomes are collected and maintained by particular universities, with only partial integration with RIS.

RIS offers a perfect departure point for policy-relevant graduate tracking studies. Indeed, it serves as a sampling base for graduate tracking surveys performed regularly by the Slovak Centre of Scientific and Technical Information (SCSTI).[13] Additionally, an annual data exchange between RIS and the RUP is performed in order to produce the graduate unemployment rate, which is also used in formulas that are applied by the Ministry of Education, Research, Development and Youth to allocate public resources to schools. Finally, RIS data on graduates is also linked with the SI database on an annual basis to produce information about the average income and employment rate of recent graduates from secondary and tertiary schools. This information is published to steer individual educational decisions of potential enrollees to secondary and tertiary schools.[14]

Additionally, linking RIS with individual student outcomes in the secondary school leaving exam revealed that the better-performing students leave to study abroad and that this outflow comprised approximately 17 percent of the cohort graduating from upper secondary programmes (Martiňák & Varsik, 2020). Administrative data, thus, documented that the dynamics of Slovak brain drain ranked among the highest in Europe and pointed out the unfavourable structure of the flows, pressuring the subsequent Slovak governments to declare this finding as being among their top policy challenges. The study is also innovative in how it combines various AD sources to identify individuals who graduate from Slovak secondary schools and continue to study abroad.[15]

Finally, the Institute of Educational Policy produced multiple policy analyses focused on practical issues in the educational systems operation that directly impact decision-making. One good example is the introduction of a regional compensation bonus in the salaries of elementary and secondary school teachers. The bonus was designed based on an analysis, using RIS data complemented by other administrative data sources,[16] documenting a higher fluctuation in regions in which teachers earn less compared to their peers in the private sector (Balberčáková & Erdelyiová Gancátová, 2022).

---

[13] https://www.cvtisr.sk/cvti-sr-vedecka-kniznica/informacie-o-skolstve/skolstvo/vysoke-skoly/uplatnenie-absolventov-vysokych-skol-na-trhu-prace.html?page_id=28928.

[14] This initiative shows a positive example of cooperation between analytical units at the different ministries, the Institute of Social Policy, and the Institute of Educational Policy. An example of such an information web can be found here: https://institutsocialnejpolitiky.sk/absolventi/

Another example using an earlier data connection of RIS to the Information System on Labour Costs can be found here: https://uplatnenie.sk/.

[15] The definition is based on the applications submitted for wasteprocessing payment, discounts for which parents are eligible if their children study at university. Another AD-based study on Slovak brain drain utilised information from health insurance payments (Haluš et al., 2017).

[16] Administrative and online data on job vacancies or the information system on labour costs.

The educational sector presents a terrain that is well covered by administrative data collection. Slovakia benefits from an integrated information system (RIS); its utilisation remains limited by the wider academic sector's access to data. In contrast, the already produced analyses show a transparent bridging to policy-making.

3.3 Health and Environmental Policy

Like the educational sector, the health sector also offers a feasible terrain to be covered by administrative data. Here, a domain-specific institution, the National Health Information Centre, is responsible for data collection and management. A series of spending reviews investigated the sector's performance, exploring the administrative data sources available and proposed specific measures. Out of the analyses with the highest societal impact, let us mention the benchmarking of prices of medical materials and equipment purchased by hospitals. In 2018, based on a Spending Review of the Health Sector (Ministry of Finance of the Slovak Republic, 2016), the Ministry of Health introduced a public procurement instruction for hospitals. The impact of its introduction between 2017 and 2018 was estimated to be approximately 40 million euro of annually saved hospital spending (Ministry of Finance of the Slovak Republic, 2022).

Another area with a wide impact on public health is environmental policy. In January 2022, Slovakia was among the first European countries to introduce a deposit-based recycling system for plastic bottles. The system was designed based on a cost-benefit model using administrative data on waste management collected by SOSR (Dráb & Slučiaková, 2018). The system was introduced smoothly, yielding a surprisingly positive impact, with the share of recycled PET bottles reaching 92 percent[17] in the second year of implementation.

In the Environmental policy domain, multiple instances of comparable analyses that propose specific policy measures that influenced particular policy decisions might be mentioned, e.g., the report on the Slovak waste management model (Slušiaková, 2023) resulted in the enhancement of the waste incineration plants and less landfilling solutions or the report on the operation of the Slovak Water-Management Company; this led to more sustainable financing and operation of the state-owned firm responsible for flood control measures (Mojzesová et al., 2023). Despite the more complex data context lacking one dominant backbone database, environmental policy offers a field in which policy analysis can translate to improvements in existing policies.

[17] Bi-annual report of the Deposit-Based Recycling System Administrator, Online: https://slovenskozalohuje.sk/wp-content/uploads/2024/03/Spravca-zalohoveho-systemu_Sprava-o-cinnosti-za-druhy-polrok-2023.pdf.

3.4 Utilisation of Administrative Data in Policy Analysis Targeting Firms

The structure of Slovak firm-level administrative data copies the usual European pattern, with the backbone being the Business Register maintained by the SOSR. Firm-level accounting data from the Registry of Financial Statements collected by the Financial Administration underlying the Ministry of Finance is publicly available online. Additionally, data on drawing public support, including the European Structural and Investment Funds or the EU Common Agricultural Policy, are collected and maintained by the institutions that distribute the support.

Agricultural production in Slovakia is highly concentrated to a limited number of large producers. This issue was highlighted by the Spending Review on Agriculture that proposed a cap on direct payments for large agricultural producers with less intensive production and employment (Ministry of Agriculture of the Slovak Republic, 2018b and 2022). The measure was supported by an ex-ante evaluation study modelling the impact of the capping on the structure of direct payments provided to the population of agricultural producers. The study was based on administrative data from the Ministry of Agriculture on agricultural producers, their production, and the subsidies received (Ministry of Agriculture of the Slovak Republic, 2018a).

Related to the core administrative database covering the whole business sector, the Registry of Financial Statements was used in testing techniques revealing fraudulent behaviour (Bačo et al., 2023), but also in designing nudging experiments aimed at increasing tax and social contribution payment discipline of small firms[18] (Priesol et al., 2024).

Despite its availability online, the Business register with the Registry of Financial Statements appears to be relatively less explored. Its policy relevance increases with linkages to additional data sources, for example, complementing it with the information on drawing EU funds or the support paid to firms during the COVID lockdowns, which resulted in three ex-post evaluation studies on the financial scheme's effectiveness (Bělín & Veselková, 2022; Mikulík & Gardoňová, 2023; Valková & Peciar, 2022).

4 Conclusions

Slovakia does not rank among the countries with long evidence-based policy-making traditions. However, the situation has started to change recently, since 2016, with the initiative seeking higher value in public spending. Extensive analytical capacities were built at the Slovak public institutions thanks to this initiative. These enabled

[18] However, the payment plance of individuals obliged to pay property tax also appeared (Rybošová et al. 2021). For an overview of behavioural interventions in the Slovak public policy, consult Sloboda, Sičáková-Beblavá and Čaplánová (2022).

an upsurge of policy-relevant analytical studies, combined with solid administrative data, collected by a broad, public bureaucracy. The newly established or enforced analytical units were a part of each responsible ministry organisationally. This proximity ensures, on the one hand, that policy relevance in the choice of the topic and, on the other, that a more direct translation of the evidence to policy-making takes place. Up to this point, the Slovak experience in enhancing the value for money in public policy can be assessed as having been successful. This can be documented by an extensive list of studies successfully translated into policy-making.

In this chapter, we outline and reference a few examples of analytical studies showing various shades of the translation of evidence into policy-making. Extracting from this experience, let us first highlight the form in which AD-based evidence was translated into policy-making. We describe an example of designing a completely new policy, such as the system of deposit-based recycling of plastic bottles. Evidence generated thanks to administrative data on waste management and retail enabled the system to set up prices, processes, and payments. Such examples of a smooth introduction of a new policy are only possible when political interest and solid analytics work in concert.

Secondly, we have also mentioned examples of evidence that served to improve the establishment of an existing policy. The regional compensation bonus in the salaries of elementary and secondary school teachers is an example of AD-based evidence that pointed to a problem, proposed a solution, and translated to an improvement of the existing formula applied to calculate teachers' salaries. In this case, the solution was politically neutral, and the evidence presented did not have to argue against vital competing interests. The case was less clear in the instance of capping direct subsidies for big agricultural producers with low-intensity production. Analogously, in this example, AD-based evidence also led to a change in the allocation formula, but with a more politically sensible result. The evidence produced had to defend the public interest against the lobbying of an economically strong interest group.

Finally, we have AD-based evidence that is policy-relevant, but that cannot be linked to either a particular policy measure or regulation. This is the evidence on the intensity and structure of the Slovak brain drain. A study, demanding in terms of data processing and analytical input, showed that it effectively frames public debate and top political priorities. Debate-framing evidence might not lead to a direct policy action, but it can still be helpful. In some cases, evidence is needed even simply to avoid an incorrect policy decision or pressing for a different allocation of resources at the top-priority (e.g. the EU) level.

We should bear the examples of impact evaluation studies on three active labour market policy programmes in mind. For the programmes supporting spatial mobility, the evidence of a positive impact on income and employment, supported by a favourable cost-benefit assessment, translated to a substantial increase in resources allocated for these programmes. A textbook example aligned with the common expectations; in the case of the training programme, the initially negative impact estimated led to a reform of the programme with a subsequent turn to a positive impact. In the case of public works, incorrectly communicated estimations of negative impact led to a break in the implementation of the programme without a suitable

replacement. Understanding the importance (and eventually also the relative undernourishment) of training in the portfolio of Slovak ALMPs blocked out the easy solution of abolishing the programme. The training example shows the importance of interpreting the evidence in a broader context. At the same time, the example of public works reminds us of the risk of over-interpreting the average treatment effects. Some communities or groups of clients suffered existential harm due to the abolition of the programme. A simpler solution was chosen because the heterogeneity behind the negative average treatment effects was not considered, and there was a lack of political will to either reform or replace the programme. As a result, some of the "good example" municipalities[19] running community firms also lost their channel for testing the employability of the long-term unemployed. The individuals locked in marginalised Roma communities lost one of the channels of getting out of long-term unemployment.

Competency combined with data access creates a necessary precondition of a better informed, but still only an interpretation of what is happening, at the implementation level or in "everyday life". Data and, to an even greater degree, the statistical models always present only a simplified representation of reality. As with any simplification, they might be helpful for better understanding, but rely on assumptions and, therefore, demand an appropriate level of humbleness in voicing our findings and policy recommendations. Policy-making in Slovakia, especially since the introduction of the Value for Money initiative, can offer plenty of examples reminding us of both the importance and limitations of AD-based evidence.

References

Act on Budgetary Rules of Public Administration No. 523/2004, Coll. [Zákon o rozpočtových pravidlách verejnej správy a o zmene a doplnení niektorých zákonov] [Online]. https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2004/523.

Act on Employment Services No. 5/2004 Coll. [Zákon o službách zamestnanosti a o doplnení a zmene niektorých zákonov] [Online]. https://www.zakonypreludi.sk/zz/2004-5.

Act on Free Access to Information No. 211/2000 Coll. [Zákon o slobodnom prístupe k informáciám a o zmene a doplnení niektorých zákonov (zákon o slobode informácií)] [Online]. https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2000/211/.

Act on State Statistics No. 540/2001, Coll. [Zákon o štátnej štatistike] [Online]. Available at: https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2001/540/.

Bačo, T., Baumohl, E., Horváth, M., & Výrost, T. (2023). Beneish model for the detection of tax manipulation: Evidence from Slovakia 1. *Ekonomický Časopis/Journal of Economics, 71*(3), 185–201. https://doi.org/10.31577/ekoncas.2023.03.01

Balberčáková M., & Erdélyiová Gancárová, M. (2022). *Možnosti regionálneho odmeňovania Pedagogických a odborných zamestnancov v školstve, Institute of Educational Policy of the Ministry of Education, Science, Research and Sport of the Slovak Republic*. https://www.mfsr.sk/files/archiv/3/regionalne-mzdy-ucitelov-komentar-final.pdf

[19] https://www.nytimes.com/2017/09/09/world/europe/slovakia-roma-spissky-hrhov-integration.html.

Bělín, M., & Veselková, M. (2022). *Job Retention Scheme in Slovakia: Impact on Dismissals and Firm Survival in the COVID-19 Pandemic, Social Policy Institute Ministry of Labor, Social Affairs and Family of the Slovak Republic*. https://www.employment.gov.sk/files/slovensky/ministerstvo/analyticke-centrum/analyticke-komentare/komentare_2022/belin-veselkova-2022-job-retention-scheme-slovakia.pdf

Bondonio, D., & Nemec, R. (2015). *Evaluation of selected structural funds and cohesion fund interventions using counterfactual impact evaluation methods: Final reports* (pp. 152–186). Bratislava: KPMG Slovensko. http://nsrr.sk/sk/hodnotenie/programove-obdobie-2007%2D%2D-2013/

Bořík, V., Ďurica, M., Molnárová, M., & Švábová, L. (2015). *Hodnotenie čistých efektov absolventskej praxe a podpory pre vykonávanie samostatnej zárobkovej činnosti (Technická správa), Ministerstvo práce, sociálnych vecí a rodiny* [Evaluation of the net effects of traineeship programmes for graduates and support for self-employment activities (Technical report)]. Ministry of Labour, Social Affairs and Family, Bratislava.

Dráb, J., & Slučiaková, S. (2018). *Skutočná cena zálohy: Analýza zavedenia systému zálohovania ednorazových nápojových obalov v SR, Institute of Environmental Policy at the Ministry of Environment of the Slovak Republic*. https://www.minzp.sk/files/iep/skutocna_cena_zalohy.pdf

Haluš, M, Hlaváč, M., Harvan, P., & Hidas, S. (2017). *Odliv mozgov po slovensky [Brain Drain in Slovak]. Institute for Financial Policy Ministry of Finance of the Slovak Republic, Komentár 2017/1*. https://www.mfsr.sk/files/archiv/priloha-stranky/4938/60/2017_1_Odliv-mozgov-po-slovensky_20170109.pdf

Harvan, P. (2011, Január 22). *Hodnotenie efektívnosti a účinnosti výdavkov na aktívne politiky trhu práce na Slovensku, Inštitút Finančnej politiky* [Evaluation of cost efficiency and of effectiveness of active policies for the Slovak labour market]. Institute for Financial Policy MF SR, Ekonomická analýza.

Hidas, S., Vaľková, K., & Harvan, P. (2016, October). *Veľa práce na úradoch práce: Efektivita a účinnosť služieb zamestnanosti. Október 2016, Inštitút Fiačnej politiky* [A lot of work at labour offices: effectiveness and efficiency of employment services]. Institute for Financial Policy, Ekonomická analýza 40.

Lubyová, M., Karasová, K., & Studená, I. (2016). Long-term effects of the traineeship programme for graduates: 4. chapter. In *Labour Market in Slovakia 2017+*. ISBN 978–80–970850-4-9.

Lubyová, M., Štefánik, M., Dováľová, G., & Karasová, K. (2014). *Analýza účinkov nástrojov aktívnej politiky trhu práce*. [Analysis of the impact of active labour market policies]. Bratislava. <https://www.researchgate.net/profile/Katarina-Moravcikova-2/publication/303228016_Analyza_ucinkov_nastrojov_aktivnej_politiky_trhu_prace_Impact_evaluation_of_interventions_of_active_labour_market_policy/links/5739831708ae9ace840d52ac/Analyza-ucinkov-nastrojov-aktivnej-politiky-trhu-prace-Impact-evaluation-of-interventions-of-active-labour-market-policy.pdf>

Lubyova, M., & Van Ours, J. (1997). Unemployment dynamics and the restructuring of the Slovak unemployment benefit system. *European Economic Review, 41*, 925–934.

Lubyova, M., & Van Ours, J. (1998). Work incentives and other effects of the transition to social assistance: Evidence from the Slovak Republic. *Empirical Economics, 23*(1/2), 121–154.

Lubyova, M., & van Ours, J. C. (1999). Effects of active labor market programs on the transition rate from unemployment into regular jobs in the Slovak Republic. *Journal of Comparative Economics, 27*(1), 90–112. https://doi.org/10.1006/jcec.1998.1571

Martinák, D., & Varsik, S. (2020) *Odliv mozgov I: Necestuj tým vlakom!, Institute of Educational Policy of the Ministry of Education, Science, Research and Sport of the Slovak Republic*. https://www.minedu.sk/data/att/bcb/18017.5534bf.pdf

Mikulík, T., & Gardoňová, K. (2023) *Prežili by aj bez pomoci. Štát mohol počas pandémie zachrániť firmy len za polovicu nákladov* (Liquidity drain I.), [They would survive also without the support. The State could have saved the firms for only half of the costs (Liquidity drain I.)]. Institute of Strategic Analyses, the Governmental Office of the Slovak Republic. https://analyzy.gov.sk/files/archiv/59/021023_ISA_Liquidity-Drain_final.pdf

Ministry of Agriculture of the Slovak Republic. (2018a). *Dopady stropovania priamych platieb na slovenské poľnohospodárstvo* [The impact of capping the direct payments in Slovak agriculture]. Diskusná štúdia IPP 2018/1, Institute of Agricultural Policy—Ministry of Agriculture of the Slovak Republic, 2018.

Ministry of Agriculture of the Slovak Republic. (2018b). *Revízia výdavkov na pôdohospodárstvo a rozvoj vidieka—Priebežná správa* [Spending Revision on Agriculture and Rural Development—Interim report], Ministry of Agriculture of the Slovak Republic, 2018.

Ministry of Agriculture of the Slovak Republic. (2022). *Revízia výdavkov na pôdohospodárstvo a rozvoj vidieka—Záverečná správa* [Spending Revision on Agriculture and Rural Development—Final report], Ministry of Agriculture of the Slovak Republic, 2022.

Ministry of Finance of the Slovak Republic. (2016). *Revízia výdavkov na Zdravotníctvo [Spending Review of the Health Sector]* https://www.mfsr.sk/files/archiv/uhp/3370/76/Finalna-sprava_revizia_ZDRAVOTNICTVO.pdf

Ministry of Finance of the Slovak Republic. (2018). *Revízia výdavkov zamestnanosti a odmeňovania vo verejnej správe: Priebežná správa* [Spending Review on Employment and Remuneration in the Public sector: Interim report]. https://www.mfsr.sk/files/archiv/5/priebezna_sprava_zamestnanost_odmenovanie.pdf

Ministry of Finance of the Slovak Republic. (2022). *Revízia výdavkov na Zdravotníctvo: Priebežná správa* [Spending Review of the Health Sector: Interim report]. https://www.mfsr.sk/files/archiv/29/Aktualizacia-revizie_zdravotnictvo_final.pdf

Mojzesová, K, Gális, M., Tóth, V., Engeľ, M., & Sokolová, M. (2023). *Po nás potopa?* [Economic and Financial Analysis of the Slovak Water-Management Company], Ekonomická analýza 17, Institute of Environmental Policy, Ministry of Environment of the Slovak Republic. https://www.minzp.sk/files/iep/analyzy/po_nas_potopa_iep.pdf

Petráš, J. (2018). *Kto chce žať, musí siať: Analýza čistých efektov opatrenia REPAS, Social Policy Institute Ministry of Labor, Social Affairs and Family of the Slovak Republic*. https://www.employment.gov.sk/files/slovensky/ministerstvo/analyticke-centrum/analyticke-komentare/kto_chce_zat_musi_siat.pdf

Petráš, J. (2019). *Dočasná práca ako šanca pre znevýhodnených: Analýza čistej účinnosti opatrenia 50j, Social Policy Institute Ministry of Labor, Social Affairs and Family of the Slovak Republic*. https://www.employment.gov.sk/files/slovensky/ministerstvo/analyticke-centrum/analyticke-komentare/analyza-50j-docasna-praca-ako-sanca-znevyhodnenych.pdf

Petráš, J. (2020). *Aktivačné práce neaktívnych neaktivujú: Analýza čistej účinnosti menších obecných služieb (MOS), Social Policy Institute Ministry of Labor, Social Affairs and Family of the Slovak Republic*. https://www.employment.gov.sk/files/slovensky/ministerstvo/analyticke-centrum/analyticke-komentare/komentare_2022/aktivacne-prace.pdf

Petráš, J. (2021). *Čo sa počas praxe naučíš, počas práce zúročíš: Analýza čistej účinnosti Príspevku na vykonávanie absolventskej praxe (opatrenie 51)—NP Absolventská prax štartuje zamestnanie, Social Policy Institute Ministry of Labor, Social Affairs and Family of the Slovak Republic*. https://www.employment.gov.sk/files/slovensky/ministerstvo/analyticke-centrum/analyticke-komentare/komentare_2022/analyza_absolventskej_praxe_final.pdf

Petráš, J. (2023). *Mentoring pomáha mladým zamestnať sa: Analýza čistej účinnosti národného projektu „Praxou k zamestnaniu, Social Policy Institute Ministry of Labor, Social Affairs and Family of the Slovak Republic*. https://www.employment.gov.sk/files/slovensky/ministerstvo/analyticke-centrum/analyticke-komentare/komentare_2023/petras_2023_mentoring.pdf

Priesol, R., Jalakšová, P., & Tužilová, B. (2024). *Odhaľ ma, ak to dokážeš: Ako motivovať živnostníkov k platbe poistného?* [Catch me, if you can: How to motivate self-employed to pay social contributions]. https://ifp.sk/odhal-ma-ak-to-dokazes-ako-motivovat-zivnostnikov-k-platbe-poistneho/

Rybošová, P., Gábik, R., Priesol, R., & Machlica, G. (2021). *Písať listy sa oplatí: ako zvýšiť daňové príjmy rýchlo a efektívne*. https://www.mfsr.sk/files/archiv/49/Komentar.pdf

Sloboda, M., Sičáková-Beblavá, E., & Čaplánová, A. (Eds.). (2022). *Behaviorálna veda v službách verejnej správy: behaviorálne intervencie v praxi* [Behavioural science in the service of the

public sector: behavioural interventions in practice], Bratislava: EKONÓM, 2022. ISBN 978–80–225-4983-7.
Slušiaková, S. (2023). *Ako von zo smetiska* [Gow to get out of the junkyard], https://analyzy.gov.sk/files/archiv/40/ako_von_zo_smetiska_iep.pdf
Štefánik, M. (2014). Estimating treatment effects of a training programme in Slovakia using propensity score matching. In *Ekonomický časopis* (Vol. 62, Issue. 6, pp. 631–645). https://www.sav.sk/journals/uploads/0620140006%2014%20%C5%A0tefanik+RS%20+F.pdf
Štefánik, M. (2018). Active Labour Market Policies in Slovakia -Overview of the Available Evidence. In *Labour Market in Slovakia 2019+*. Bratislava (pp. 38–55. ISBN 978–80–7144-296-7). http://ekonom.sav.sk/uploads/journals/390_labour_market_in_slovakia_2019.pdf
Stefanik, M. (2021). Shifting the training choice decision to the jobseeker—The Slovak experience. *Labour, 35*(2), 192–213. https://doi.org/10.1111/LABR.12189
Štefánik, M., & Karasová, K. (2016). Karasová, Katarína. Evaluating the Effectiveness of Policy Measures to Support Spatial Mobility in Slovakia: the Contribution for Commuting to Work and the Contribution for Resettlement. In *Ekonomický časopis* (Vol. 64, Issue. 8, p. 711–736).
Štefánik, M., Karasová, K., & Studená, I. (2018). Can supporting workplace insertions of unemployed recent graduates improve their long-term employability? Evidence on the treatment effects of the contribution for the graduate practice in Slovakia. *Empirica.* https://doi.org/10.1007/s10663-018-9413-y
Štefánik, M., & Lafférs, L. (2024). Supporting the right workplace experience: A dynamic evaluation of three activation programmes for young job seekers in Slovakia. *Journal of Labour Market Research.* (In Print).
Svabova, L., Kramarova, K., & Gabrikova, B. (2022). Counterfactual assessment of the allowance for school-leaver practice performance as a measure of. *Active Labour Market Policy in Slovakia, 1*, 99–123. https://doi.org/10.31577/ekoncas.2022.02.01
Vaľková K., & Peciar V. (2022). Zásady poskytovania Prvej pomoci (na trhu práce). *[Rules of providing the First-aid (at the labour market)] Institute for Financial Policy, Ekonomická analýza 40*. /Publikácie IFP/Ekonomické analýzy. https://analyzy.gov.sk/files/archiv/98/25935_material_hlasovanie.pdf
Vaňo, B. (2014). Sčítanie obyvateľov na slovensku—súčasnosť a perspektívy (Population census in slovakia—present situation and perspectives). *Slovak Statistics and Demography, 24*, https://ssad.statistics.sk/SSaD/wp-content/files/3_2014/3_2014_clanok_9_Vano.pdf

Miroslav Štefánik currently serves as the director of the Institute of Economic Research of the Slovak Academy of Sciences. His educational background combines sociology and economics with quantitative methodology, underlined by solid data processing skills. His research focuses on processing big data to produce policy-relevant information, such as program impact evaluations and forecasting. Among the topics covered are active labour market and educational policies. He regularly publishes in peer-reviewed international scientific journals and actively comments and contributes to creating strategic policy documents.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 26
Unlocking Slovenia's Administrative Data: Abundant Sources, Complex Access



Matija Vodopivec

Abstract Administrative data in Slovenia hold significant potential for research across many academic and policy areas. Slovenia's Statistical Office, SURS, serves as the entry point for accessing the substantial data available in Slovenia, enabling access to the data that it holds in-house as well as accessing data from other administrative sources. While an established formal procedure exists in theory, accessing and linking these data sources can entail navigating a complex landscape of approvals and arranging technical access in practice. Researchers must plan their data requests carefully, given that institutional willingness and technical capacity greatly influence both the ease and timeliness of access. Collaboration with data-holding institutions can also help to streamline the process. Despite the time-consuming nature of securing access, particularly with external datasets, the wealth of administrative data available presents substantial opportunities to gain impactful research and policy insights.

1 Introduction

Access to individual-level administrative data for research purposes has a fairly long tradition in Slovenia. In the late 1980s, for example, researchers used firm-level data to examine the question of redistributions across firms in the absence of subsidies in socialist Yugoslavia (Kraft & Vodopivec, 1992). Slovenia's central bank, Banka Slovenije, also used its rich data to conduct various analyses, such as an analysis of the banking network in the early 1990s (Repanšek, 1994). This was subsequently expanded to hundreds of studies in the ensuing decades, spanning a wide variety of different disciplines and types of data (SURS, 2024a, 2024b, 2024c).

External researchers may access confidential administrative microdata for research purposes in Slovenia through three government entities. The primary source for most researchers is the Statistical Office of the Republic of Slovenia (SURS), which has regular access to a large number of administrative data sources and its own

M. Vodopivec (✉)
OECD, Paris, France
e-mail: Matija.Vodopivec@oecd.org

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_26

survey data.[1] Researchers can, thus, access the databases that SURS uses to generate the statistics that it publishes. When accessing microdata through SURS, researchers also have the option of requesting additional administrative data from other sources that SURS does not use for statistical purposes, which SURS pseudonymises before making it available to researchers. In addition to the Statistical Office, researchers may also access administrative microdata through both the National Institute of Public Health (NIJZ) and the Slovene central bank, Banka Slovenije (BS). In contrast to data available from SURS—which, in principle, can bring in administrative data from many external sources for linking—the possibilities for linking external data sources at NIJZ and BS are more limited. Unless researchers are specifically looking for specialised databases maintained by NIJZ or BS, the typical course of action is to request data access through SURS. Given that accessing data through SURS is likely to be the most relevant channel for most users, this is also the focus of this chapter.

The structure of this chapter is as follows. The chapter first discusses the process of receiving access to data through SURS, detailing the necessary administrative steps. It then discusses which databases can be accessed, based on both data that is already available in-house at the Statistical Office as well as that from external sources. This is followed by a discussion of how the accessing of data and results through the Statistical Office occurs in practice. The chapter concludes with a discussion of future directions that administrative data access might take in Slovenia.

2 Process of Receiving Access to Data Through the Statistical Office (SURS)

The process of acquiring access to microdata through SURS proceeds in several steps, with SURS asserting that the procedure is broadly in line with the best practices suggested by UNECE (2007). The process proceeds in the following steps:

1. **Ascertaining the data to be requested by taking data availability into account.** This first step involves liaising with SURS staff to determine precisely which data are available in-house through SURS and which data would need to be requested from external administrative sources. SURS has access to a rich set of data that it uses for its statistical reporting purposes and, in principle, the underlying data could be made available to researchers (see section below for details). If the researcher would like to request additional administrative data from other sources—including, possibly, coverage of time periods for data that are not available through SURS—then the researcher needs to determine whether the administrative entity who manages the database would agree to sharing the data with SURS. The researcher is required to have detailed knowledge of precisely

[1] To facilitate the subsequent work of researchers who may be interested in conducting research using Slovene data, the key institutions in this chapter are referred to according to their Slovene acronyms.

which data they would like to request—down to the level of individual variables requested and exact time coverage—before proceeding to the next step.

2. **Submitting an application to the SURS Committee for Researchers**. This advisory body to the SURS Director-General processes the applications (SURS, 2024a, 2024b, 2024c). The application must include the purpose of the research, a short summary of the analysis, the precise data sources requested and the exact list of variables being requested from each data source. It must also indicate the software foreseen for the data analysis (this is relevant mainly if the researchers intend to use the existing computers available to researchers, not their own hardware/software—see discussion below). The committee typically reaches its decision within several weeks.
3. **Signing the legal contracts and documents to obtain access to the data, assuming the application is successful.** This includes a contract between the researcher's employer, SURS and any external entities providing data. It also involves signing person-specific confidentiality declarations.

Access to microdata through the Statistical Office is available to individuals from academic and research institutions. This process is facilitated if individuals have a research ID from the Slovene Research and Innovation Agency (ARIS).[2] However, researchers from abroad without an ARIS research ID can also qualify. In the absence of having an ARIS research ID, it can be useful to partner with either a local researcher or institution. Having a local research partner can also help at other steps of the process.

This willingness to share data can vary considerably across institutions and over time. In all cases, the legal basis for such sharing is set out in Article 69 of the Slovenian Personal Data Protection Act (ZVOP-2), which specifies that "Research organizations and researchers who are bound by ethical principles in their research and methodology… may, for the purposes of the research, obtain from the controller personal data, including special categories of personal data" (Official Gazette of the Republic of Slovenia, 2024a, 2024b).

The duration of the process to acquire legal access can vary considerably in practice, but takes a minimum of two months. The involvement of outside institutions, without prior experience in providing microdata to SURS, can prolong the process considerably, while the procedure is typically quicker with institutions that have established the precedent of sending such data. However, it can also vary over time within a given institution, depending on their technical capacity to provide the data as well as the staff's willingness to cooperate in the process.

Access to data provided through SURS is free of charge to users—SURS provides these services for researchers as a public service. In exchange for the free access, researchers using data accessed through SURS are required to cite SURS as the source of their data in their articles, presentations, posters and in other materials. In addition, researchers are required to notify SURS about publications arising from

[2] Conditions for receiving a research ID from ARIS are available on the ARIS website—more information is available at https://www.arrs.si/sl/evidreg/evid/gradivo/obrazci/23/podatki-RO-2023.asp.

the use of microdata, which SURS uses to track the impact of these works (the list of publications is available online—see SURS (2024a, 2024b, 2024c).

The procedure for accessing the data is the same irrespective of the sensitivity of data for which access is being requested. However, assuming that access is granted, the sensitivity of the data may result in either different conditions of access (e.g., only in-person access) or different stringency concerning the implementation of the data minimisation criteria (the criteria that only data that are absolutely necessary for the analysis are to be provided). For example, accessing health data may require severely paring down the sample of individuals for whom data are requested or the categories of illnesses that are to be included in the requested data.

3 Databases Available Through the Statistical Office

SURS can, in principle, grant researchers access to all of the data that it collects for the purposes of generating statistics. These are detailed in the annual programme of statistical research (LPSR). SURS also facilitates access to data provided by external entities, thereby serving as an intermediary which receives the raw data with the personal (or firm-level) identifiers that it then replaces with pseudonyms and provides to researchers in a secure environment.

The data provided to researchers by SURS also contains the metadata necessary for analysis. The metadata is provided in both Slovene and English (upon request). The quality of the metadata and accompanying documentation can vary considerably for outside sources. The quality is generally excellent for data provided by SURS. Additional supporting documentation on the data provided by SURS is publicly available in the methodological, supplemental information to the publicly available statistics in its databases.

3.1 Data that is Already Available In-House at the Statistical Office

As mentioned previously, the databases used by SURS for its statistical reporting purposes are also available to researchers. The precise data—encompassing both administrative data and that obtained through surveys administered through the Statistical Office—is updated annually and is published as a programme of work. Table 26.1 summarises the types of data contained in the programme of work for 2024.

Some of the topics above are partially covered by two other entities, NIJZ and BS. These relate to some of the databases that cover health and finance/banking, respectively. In such cases, they are the relevant institutions to contact regarding microdata. In principle, these institutions may also transfer microdata to SURS that

Table 26.1 A wide array of different data are obtained for statistical purposes in Slovenia. Data included in the 2024 programme of work for statistical research

	Type of data	Providing institutions	Examples of data provided
1	National accounts	Ministry of the interior, agency of the Republic of Slovenia for public legal records and related services (AJPES), ministry of finance	GDP data, employment, business and financial statements, tax records, balance of payments and electricity usage
2	Regional accounts	AJPES, ministry of finance	Regional GDP, employment statistics by region and business activity by sector and region
3	Prices	Retailers, pharmacies, ministry of infrastructure	Consumer price index, housing and rental prices, energy prices and pharmaceutical product prices
4	Population	Ministry of the interior, surveying and mapping authority of the Republic of Slovenia (GURS)	Demographics, births, deaths, marriages, divorces and migration statistics
5	Labour market	Pension and disability insurance institute of Slovenia, health insurance institute of Slovenia, AJPES	Employment and unemployment data, labour costs and occupational statistics
6	Living standards	AJPES, ministry of finance, health insurance institute of Slovenia	Household income, expenditures, housing conditions and poverty indicators
7	Education	Ministry of education, science and sport, national examinations centre	School enrolment, graduation rates, adult education, vocational training statistics
8	Culture	Ministry of culture, libraries, museums	Participation in cultural activities, number of cultural institutions and employment in cultural sectors
9	Health and social security	National institute of public health, health insurance institute of Slovenia, ministry of labour, family, social affairs and equal opportunities	Health insurance coverage, social welfare benefits and hospital data
10	Crime and safety	Ministry of justice, ministry of the interior	Crime statistics, convictions, criminal offenses and road safety

(continued)

Table 26.1 (continued)

	Type of data	Providing institutions	Examples of data provided
11	Business entities	AJPES, ministry of economic development and technology	Number of businesses, legal structure, ownership and financial performance
12	International economic relations	Bank of Slovenia, ministry of finance, customs administration	Foreign trade, foreign investments and balance of payments
13	Research and development, science and technology	Ministry of education, science and sport, research institutions	R&D spending, innovation statistics, patents and workforce in science and technology
14	Mining and manufacturing	Mining companies, manufacturing businesses	Production levels, employment and financial performance in mining and manufacturing sectors
15	Energy	Ministry of infrastructure, energy suppliers	Energy production and consumption, electricity and gas prices and renewable energy usage
16	Construction	Ministry of Environment and Spatial Planning, construction companies	Building permits, construction costs and housing and commercial construction statistics
17	Trade and services	Retailers, ministry of economic development and technology	Sales, service industry data and employment and wages in trade
18	Financial intermediation	Bank of Slovenia, commercial banks, insurance companies	Financial accounts, loans, deposits and insurance claims and premiums
19	Tourism	Ministry of economic development and technology, tourist accommodations	Visitor numbers, overnight stays, tourism revenues and types of accommodations
20	Transport and communications	Ministry of Infrastructure, telecommunications providers	Freight and passenger traffic, communication services data and internet and mobile usage
21	Agriculture, forestry and fisheries	Ministry of agriculture, forestry and food, Slovenian forest service	Agricultural production, forestry activities and fishing yields
22	Environment	Ministry of environment and spatial planning, environmental agency	Air and water quality, waste management and environmental protection statistics

(continued)

Table 26.1 (continued)

	Type of data	Providing institutions	Examples of data provided
23	Cross-sectoral indicator systems	Various government ministries and agencies	Cross-sectoral indicators for socio-economic planning and development goals
24	Registers	Ministry of the interior, surveying and mapping authority, AJPES	Business register, property and land ownership records and address and population registers
25	Miscellaneous	Various government bodies	Other data that does not fall under specific categories, often supporting EU reporting requirements

Notes In most cases, data are to be provided at a granular (not aggregated) level. For example, labour market statistics are to be furnished at the level of the individual and financial statistics are to be provided at the level of the firm
Source Official Gazette of the Republic of Slovenia (2024a, 2024b).

will be combined with other administrative data sources. In contrast to SURS, the other entities may also charge users a fee to access their data.

3.2 Examples of Additional Administrative Databases that can in Principle be Linked

In addition to the databases SURS has in-house, researchers can request access to databases owned by other government entities. This includes publicly available databases and those that are not publicly available. While researchers can certainly access publicly available databases without the involvement of SURS, SURS can enable such databases to be linked with non-publicly available databases. A list of these databases is available at the Open Data portal which is maintained by the Ministry of Public Administration (Ministry of Public Administration, 2024).

Selected and non-exhaustive examples of such data include the following:

- **Publicly available geodetic data** (https://www.e-prostor.gov.si/en/access-to-geodetic-data/). The data from the national topographic and coordinate systems include remote sensing information, topographic models, national topographic and general maps, a digital elevation model and geographical names. It also encompasses data from the national coordinate system, national border records and registers of spatial units and addresses. Public infrastructure data includes the public infrastructure cadastre and network termination points. Land parcels, buildings and other real estate cadastre data are also included.

- **Data from the Slovenian Meteorological Agency**. The collection includes data from meteorological stations and other measurements. These include current and past weather, cloud cover and type, cloud base height, visibility, crop conditions and soil conditions. Measurements include wind direction and speed, air temperature, extreme air temperatures (highest and lowest) and soil temperature at various depths.
- **Data on jobseekers registered with the public employment service**. This includes information about jobseekers' receipt of unemployment insurance benefits, detailed social transfers by category and participation in active labour market programmes.

3.3 Linking Data Generated by Researchers

A final category of data that can be linked to administrative sources, using SURS as an intermediary, is from researchers' own data. This can be useful, for example, to examine the effectiveness of a trial intervention where information on participation in an intervention is applied to the administrative data, with the administrative data offering a rich basis for generating control groups and calculating the intervention's net effects. It may also be useful to complement the administrative data with survey data so as to incorporate information that is specifically relevant to the researcher.

For researchers wishing to conduct an ad hoc survey that they intend to complement with administrative data, SURS can provide the researchers with some individual-level (personal) data necessary for survey design. They can generate representative samples of individuals with appropriate weights, including the information necessary for contacting the individuals. This includes personal data, such as the name and surname of the individual, their residence, year of birth and gender. SURS can also help generate representative samples of firms.

Prior to undertaking any research under the presumption that it will be possible to link external sources with other administrative data via SURS, researchers should consult with SURS in advance to definitively ascertain whether this will be possible. Ideally, this would include submitting an application through SURS' Committee for Researchers, based on the procedure described above *before* undertaking the survey. This may help to clarify any necessary steps that need to be taken in order to enable the data to be linked. For example, if the researcher intends to link survey data with administrative data, then the survey should explicitly request the consent of the individual or entity being surveyed, informing them of the data that will be linked. This consent is necessary to establish a legal basis for the survey responses to be linked with data from administrative sources.

4 Methods of Accessing Data and Results Through the Statistical Office

Once the necessary contractual agreements for accessing data through SURS have been established, the next step is for SURS to prepare the data files for analysis. This includes obtaining any necessary data from external entities that have been transferred to SURS, preparing the supporting documentation and replacing the personalised identifiers with pseudonyms. The analysis of the microdata is conducted on a secure computer that is located on the premises of SURS. This can be conducted either in-person, through the so-called secure room at SURS, or through remote access.

Researchers have the option of conducting the analysis of the microdata either on computers provided by SURS or on their own computers (i.e., ones that they hand over to SURS for this purpose). The computers provided by SURS have the benefit of being readily available, but the drawback is that they may not have the technical capabilities for analysing large databases. In addition, the computers provided by SURS are used by multiple researchers, requiring additional coordination to ensure that they will be available for use at the desired time slots. If the researchers provide their own computer, then SURS will use technical means to ensure that data cannot be transferred by researchers (e.g., through locking the USB ports and by disabling internet access). Having a dedicated computer in the SURS secure room has several benefits, but it also has certain drawbacks—for example, subsequently installing additional software on the computers once they have been delivered to SURS is very difficult in practice.

In order to obtain any results from analyses using microdata outside of the secure environment, the results must undergo thorough checks by SURS staff to ensure that personal data are not being revealed. This process is relatively straightforward in the case of regressions, for example, where sufficient sample sizes are generally interpreted as maintaining confidentiality. Frequency tables, conversely, must undergo exhaustive checks, a time-consuming process for both researchers and SURS staff. (Note that this procedure is only relevant for statistics or results that researchers would like to obtain for publication; results that remain on the secure room computers are not subject to this level of scrutiny or procedures.)

In addition to accessing confidential microdata through the procedures described above, researchers can also access a subset of the databases described in the previous section that SURS makes available through the SOVD large file repository. This provides researchers with statistically protected microdata, which are the results of specific survey statistical research that is conducted on small samples. SURS also enables researchers to use statistically unprotected aggregated data.

SURS has an advisory body comprised of researchers who can be a useful source of additional information (SURS, 2024a, 2024b, 2024c). This so-called "Council of Researchers" is intended to exchange experiences, to make proposals and to facilitate the flow of information between SURS and institutions that use microdata.

5 Directions for Future Development of Administrative Data Access in Slovenia

As this chapter has highlighted, Slovenia's administrative data holds considerable potential for researchers. The country's datasets span a wide array of domains, including social, economic, health and environmental data. This extensive array of data provides researchers with a useful resource through which to explore a diverse set of topics such as labour market dynamics and public health to climate science. By linking these datasets, researchers can generate insights that are not only relevant to national policy, but that also contribute to our broader scientific understanding.

However, several limitations are worth emphasising here. Despite the vast amount of data at the researchers' disposal, accessing and linking administrative datasets often remains a cumbersome process in practice. Depending on the data requested, the procedural hurdles involved in obtaining the necessary approvals from multiple agencies and ensuring compliance with the regulations as they are currently enforced can sometimes result in significant delays. Moreover, the institutional willingness to share data can vary across agencies, with some institutions having more experience in providing microdata than others. Additionally, the technical capacity of these entities to handle data requests can be uneven, thereby further complicating access for researchers.

Several potential solutions should be considered in light of these limitations. First, the fact that there is a legal basis for administrative data to be shared with SURS should be made explicit by the relevant authorities that are tasked with enforcing data protection legislation and by Slovenia's information commissioner in particular. This could help to clarify that Slovenia's implementation of GDPR, together with the additional legislation such as the law outlining the mandate of SURS, collectively allow for such data transfers, assuming that certain conditions have been met. It would be helpful to make this clear with specific examples. Second, SURS could take additional steps to help to ease the use of data by researchers. While the data documentation (metadata) provided alongside the raw data is generally excellent, coding the values into analysis-ready formats still entails considerable additional work for researchers (moreover, this work is duplicated across researchers). This additional step could be avoided if SURS provided data in analysis-ready data formats (for its most commonly used datasets at least). This could follow the practice employed by other statistical agencies, such as Eurostat, which provides individual-level data from its surveys to researchers in several formats. Finally, additional staffing resources could be devoted to the team within SURS that is tasked with working with researchers to help facilitate their work. Despite its limited resources, SURS has made substantial strides in ensuring data accessibility and confidentiality for researchers. With further support, SURS could continue to expand its capacity to serve the research community even more effectively.

References

Kraft, E., & Vodopivec, M. (1992). How soft is the budget constraint for Yugoslav firms? *Journal of Comparative Economics, 16*(3), 432–455. https://doi.org/10.1016/0147-5967(92)90166-5

Ministry of Public Administration. (2024). *Odprti Podatki SLovenije [Open Data Slovenia]*. Retrieved August 1, 2024, from https://podatki.gov.si/dataset/

Official Gazette of the Republic of Slovenia. (2024). *Letni program statističnih raziskovanj za 2024 [Annual statistical program of work for 2024]*. https://pisrs.si/pregledPredpisa?id=DRUG5249

Official Gazette of the Republic of Slovenia. (2024). *Zakon o varstvu osebnih podatkov (ZVOP-2) [Personal Data Protection Act]*. https://pisrs.si/pregledPredpisa?id=ZAKO7959

Repanšek, B. (1994). "Regional analysis of banking operations in Slovenia [Regionalna analiza poslovanja bank v Republiki Sloveniji]", *Prikazi in analize*, Vol. II/4.

SURS. (2024a). *Izbrani primeri raziskovalnih analiz na mikropodatkih državne statistike [Selected examples of exploratory analyses using micro-data from national statistics]*. https://www.stat.si/StatWeb/File/DocSysFile/13704/OBJAVE_RAZISKOVALCEV_objavljene_do_2023_SLO.pdf

SURS. (2024b). *Sosvet za sodelovanje z raziskovalci [Co-council for collaboration with researchers]*. Retrieved August 1, 2024, from https://www.stat.si/statweb/NationalStatistics/AdvCommitteesDescription/98

SURS. (2024c). *Statistical confidentiality: Fundamental principles*. Retrieved July 25, 2024, from https://www.stat.si/StatWeb/en/FundamentalPrinciples/StatConf

UNECE. (2007). *Managing statistical confidentiality & microdata access: Principles and guidelines of good practice*. United Nations. https://unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

Matija Vodopivec is a senior economist in the OECD's Directorate for Employment, Labour and Social Affairs. His work focuses on supporting governments to design and implement evidence-based policies in the field of employment services. Prior to joining the OECD, he worked in the analysis division of the Slovenian central bank and worked in economic consulting in Boston (USA). He holds a Ph.D. in economics from the University of Ljubljana (Slovenia) and a Bachelor of Arts from Macalester College (Minnesota, USA).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 27
State of the Art in Access to Administrative Data in Spain



Josep Amer-Mestre and Manuel Serrano-Alarcón

Abstract This chapter describes the state of the art in terms of access to administrative data in Spain. It describes microdata sources from administrative records that could be leveraged for research and policy evaluation on policy areas such as the labour market, healthcare, education, and tax policy. We also provide examples of existing studies that have used these datasets to produce evidence about policy effectiveness. Suggestions are provided about strategies by which to both improve and expand access and linkage of administrative records for research purposes.

1 Introduction

In this chapter, we outline the relevant characteristics of the main databases that have individual-level information that has been constructed on the basis of both national and regional administrative records in Spain. This chapter's scope is to describe the main features of the different sources of microdata that can be leveraged to produce evidence about policy effectiveness in different policy domains. The microdata sources covered in this chapter have been selected based on their data quality, degree of accessibility, and policy relevance.

Data access applications are generally studied on a case-by-case basis by each institution, even though there are relevant differences in terms of the accessibility of the databases covered. Applications are generally free of charge and open to the research community, including the international community. The application's language and most of the data documentation are in Spanish, although some data holders have started to offer their documentation in English too. Along with the most relevant features of each dataset, we have also included examples of academic studies that used the data covered in these databases to produce evidence that might inform

J. Amer-Mestre · M. Serrano-Alarcón (✉)
European Commission Joint Research Centre, Ispra, Italy
e-mail: manuel.serrano-alarcon@ec.europa.eu

J. Amer-Mestre
e-mail: josep.amer-mestre@ec.europa.eu

© The European Union 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_27

policy evaluation. The content of this chapter can be useful to researchers in social sciences, data analysts, and policymakers with an interest in policy evaluation.

Spain is currently not leveraging the full potential of their existing administrative data for public policy evaluation and for evidence-based policymaking (AIREF, 2020), as detailed by the Spanish Fiscal Independent Authority (AIREF). Some of the barriers to both the access and use of administrative data for policy evaluation identified in the report include a lack of common protocols for data exchange and sharing, lack of technical and human resources for data management and sharing, a perceived patrimonial attitude of some institutions regarding the data in their possession, dispersion of the data across different institutions and, overall, a large difficulty with linking data across institutions.

However, some progress has been made in the last few years. Some institutions, such as the Social Security Administration, the Tax Agency, and the National Institute of Statistics (INE) have developed protocols to apply and access data from their records for research purposes independently. Moreover, each of these institutions has also developed its procedures so as to guarantee the confidentiality of personal data and for publishing results (AIREF, 2020). Applications and data access requests are normally studied on a case-by-case basis by each institution based on such protocols.

There are also relevant examples where the use of administrative data for impact evaluation of public policies has taken off at the regional level in the last few years. Spain is a highly decentralised country in which many relevant policies (such as healthcare and education) depend on the autonomous regions (i.e., *Comunidades Autonomas*). Thus, regions are also in charge of collecting and administering data on personal records that can potentially be used for research and policy evaluation.

The remainder of this chapter is organised as follows: Sect. 2 covers the sources of micro data with labour market information; Sect. 3 concerns Health and Demography; Sect. 4 covers Education and Training, while Sect. 5 is devoted to investigating Income, Wealth, and Firm-level data. Section 6 both discusses a new initiative which involves merging different microdata sources and concludes the piece.

2 Labour Market

The Social Security Administration was one of the first institutions to provide researchers with structured datasets containing information from their vast records, covering different aspects of the individual's experiences in the labour market. Recently, a new internet platform has consolidated all of the relevant information and has streamlined the application process for available microdata files generated from Social Security records.[1] Researchers can apply for its use for non-commercial

[1] Link to the data platform: https://portaldatos.seg-social.gob.es/home. A brief description of the databases is generally available in English on the website, although the core data documents are provided only in Spanish.

purposes free of charge. Below is a selection of the most relevant microdata files that the Social Security Administration makes available for researchers:

– **Sample of Employment Histories**[2]: Contains a 4% representative sample of all workers and pensioners who registered and appeared in the Social Security records in the reference year. The reference population, from which the sample is drawn, contains all employed and self-employed workers, unemployment beneficiaries, and pensioners. However, it does not contain unemployed individuals not receiving unemployment benefits, individuals registered with the Social Security only for healthcare assistance, or recipients of pensions or benefits not administered by the Social Security. All of their records are added to the dataset in order to construct each selected individual's employment history. The sample's longitudinal design ensures that the same individuals are kept across subsequent years unless they stop being registered with the Social Security. Sampled individuals' information is expanded by linking them with the Residents Registry, and with information on their yearly labour earnings from each employer.

  The first version of the sample of employment histories dates back to 2004. The minimum level of observation provided is at the employment spell level with information about its duration, contract type, part-time, sector of activity, employer identifier, workplace location, and social security contributions. This information is available for each employment spell since the individual first registered with the Social Security in 1967. Information on the type of contract only starts to be reliable from 1997 onwards. Individual characteristics about the individual sampled, such as age, gender, place of residence, and nationality are also included. Below we cite some examples of research performed on different labour topics that have used this data in their analysis: Roca and Puga (2017) on the determinants of higher earnings in cities; De Quinto et al. (2021) on the gender earnings gap after childbirth; Güell and Lafuente (2022) on the determinants of unemployment duration; and Garcia-Louzano et al. (2023) on short-term work arrangements and human capital accumulation and wages.
– **Employer-Employee Panel**[3]: Similar to the sample of employment histories, this dataset contains a sample of employees and their employment histories. The main difference between the two has to do with the sampling unit, which in this case is the employer plant rather than the individual worker. By design, only employed individuals are included, while those in any different situation, such as self-employment, receiving unemployment benefits, or any other pension or benefit are not. Moreover, employers using special Social Security schemes to hire their employees, such as artists, fishermen, miners, farmers, housework personnel, football players, etc., are also not considered as part of the reference population.

  The dataset includes the same information that is contained in the sample of employment histories for all of the employees working at the plant sampled, excluding the information stemming from the Residence Register and the Tax

[2] Known as *Muestra Continua de Vidas Laborables* or MCVL in Spanish.

[3] Known as *Panel de datos de empresas/trabajadores* or PET in Spanish.

Records. This data allows the user to observe the complete working histories of all employees who were working in the plants sampled in the reference year. Employer plants are first stratified according to the number of employees and their identifiers are selected in proportion to their sector of activity in each stratum. Two panels, which cover different years, were constructed, one covering the period 2013–2016 and the other 2017–2020. Each panel uses a different sample of plants; therefore, a plant that was drawn in one panel does not necessarily appear in the other. A refreshment sample is drawn from each stratum in order to substitute the plants that disappeared for each subsequent year covered in the panel.

– **Sick Leaves and Work Accidents**: Research devoted to understanding the impacts of both labour and healthcare policies on the health and well-being of individuals has relied upon the information gathered through work leaves, due to health-related reasons. In Spain, the information on sick leaves can be obtained through two different datasets that are administered by Social Security, depending on their nature of it (i.e., whether temporary or permanent).[4] These datasets contain information on all of the sick leave spells granted and include details on the leave itself, such as length, a medical diagnosis, whether it is due to a relapse, as well as on basic characteristics of the beneficiary, the employer and their labour relationship. The information contained in either dataset allows the researcher to identify whether the leave spell was due to a common sick leave reason, a work accident, or because of an occupational disease.

For example, this data is used in Marie and Vall-Castelló (2023) in order to investigate the impact of a massive reduction in paid sick leave benefits on worker absence. Social Security, through its platform for researchers, offers the possibility of linking the information about sick leave spells with the employment information contained in their records, such as those in the sample of employment histories (MCVL) described above. Such linked data is used in Holub et al. (2021) to study the impact of air pollution on labour supply. The Ministry of Employment administers a separate dataset with very detailed information about each work accident recorded.

The Social Security Administration offers the possibility to link some of their different datasets, and to access these and other higher-level security datasets, through their secure data rooms or via a remote server. Over the past two decades, the Spanish Social Security Administration has made significant efforts to make the information contained in their administrative records available to researchers. The creation of an online platform with detailed information about all the available datasets, and clear information on their contents and how to apply to access each of them, is the latest evidence of its commitment to continuing with this mission. We believe that the next challenge for the Social Security Administration is to work towards facilitating the linkage of their datasets with administrative records within other branches of both the central and regional administrations. If this agenda is achieved, then it will broaden

[4] They are known as "*Altas de incapacidad temporal*" and "*Altas de incapacidad permanente*" in Spanish, respectively, and are administered by the National Institute of Social Security (*Instituto Nacional de la Seguridad Social*).

research into the effects of public policies on other areas that are currently limited, due to reliance on disconnected datasets.

3 Health and Demography

Various databases offer granular information that can be used to understand both the population's demographics and the health status and healthcare use thereof. We have included some notable examples below:

Hospital Morbidity Survey.[5] This data is held by the National Statistics Institute (INE) and provides information on nearly all hospital discharges in Spain. Its aim is to measure morbidity according to the discharge diagnosis. Data is collected through a hospital information system, called the Basic Minimum Set of Data (Conjunto Minimo Basico de Datos, CMBD), which includes homogenous administrative and clinical data collected when the patient is discharged. In particular, the dataset contains information on diagnoses (following the International Classification of Diseases, ICD), procedures, dates of admission and discharge, province of residence, and the patient's sex and age for each hospital discharge. While it does not contain individual identifiers that would allow for individuals' healthcare use patterns to be followed over time, it does offer valuable information for analysing trends in healthcare utilisation and disease prevalence. Hospital discharge data may be used both to evaluate healthcare policies and any and all public policies that may affect the health of citizens. For instance, Garcia-Pérez et al. (2024) used this data to study the effect of receiving a long-term unemployment subsidy on workers' health and found that this led to a reduction in injury hospitalisations. A more tailored extraction of CMBD data, including patient identifiers and extended information on clinical procedures, costs, and hospital characteristics, can be performed following a dedicated procedure.[6]

Death Statistics According to Cause of Death. This database, from the National Statistics Institute, includes the total volume of all deaths that occurred in Spain each year. Each entry in the dataset includes the cause of death, date of death, region of residence, gender, age, and other socioeconomic variables, such as education level or marital status.

Although the data lacks an individual identifier, it can still be used to conduct valuable analysis with public policy implications. For instance, it has been used to study the effect of a reform that was implemented in Spain that restricted access to public healthcare to undocumented immigrants (Juanmarti Mestres et al., 2021). The study

[5] All microdata from the National Statistics Institute (INE) can be accessed at Productos y servicios / Información / Ficheros de microdatos (ine.es). A brief description of the databases is generally available in English on the INE's website, although the core data documents (such as questionnaires and codebooks) are only available in Spanish. Access to INE microdata is open to the research community and is free of charge.

[6] An extraction of CMBD data can requested at: https://www.sanidad.gob.es/estadEstudios/estadisticas/estadisticas/estMinisterio/SolicitudCMBD.htm.

shows that this reform led to an increased mortality of undocumented immigrants. González and Rodríguez-González (2021) aggregated this data at the municipality level in order to study the evolution of gender and socioeconomic inequality in mortality over time. They found decreases in mortality rates across all age groups, particularly among men, between 1990 and 2018. However, they also observed an increase in inequality among older men, something that was attributed to greater improvements in cancer mortality among men residing in wealthier areas.

Birth Certificate Data. This dataset includes the sum total of births registered annually in Spain. For each birth, it includes the date of birth, newborn health measures such as whether or not it was a premature birth, birth weight or mortality, and various sociodemographic variables about the family such as family composition, country of birth, place of residence, civil status, or education level.

This database has been used to study the effect of several family policies concerning fertility. For instance, Farré and González (2019) studied whether the introduction of a two-week paid paternity leave affected fertility. They found that fathers who took this leave were less likely to have an additional child in the following years, suggesting that the policy actually led to a postponement of fertility for subsequent children. Using this data too, González and Trommlerová (2023) found that the introduction of a one-time universal child benefit of €2,500 euros in 2007 led to a 3 per cent increase in birth rates, whereas its subsequent cancellation in 2010 lead to a larger drop of 6 per cent in birth rates.

Primary Care Drugs Prescription Database (*Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria*, BIFAP)[7] provides access to anonymised individual-level clinical data from various Spanish regions covering 17% of the Spanish patients who attend primary care. BIFAP data is collected by primary care physicians and is managed by the Spanish Agency of Medicines. BIFAP contains data about drug prescriptions, health conditions, and procedures, as well as some basic sociodemographic variables. Researchers can request tailored data extractions from BIFAP, thereby ensuring patient anonymity, on condition that their research proposals have been approved by the BIFAP scientific committee. This database serves as a valuable resource for conducting research in pharmacoepidemiology, drug safety, healthcare utilisation patterns, and for evaluating the health effects of public policies on the population. For instance, Borra et al. (2021) investigated whether the universal child benefit, explained previously, improved the health of children later on in life. They found no effect on some of the outcomes available in BIFAP, such as prescriptions and primary care visits.

[7] Remote access to BIFAP data can be requested by submitting a research proposal to BIFAP Base de datos para la Investigación Farmacoepidemiológica en Atención Primaria. Basic information about the database can be found in English, while the research proposal is submitted in Spanish.

3.1 Regional Databases

The healthcare system in Spain is decentralised and is managed by the 17 autonomous regions at the regional level. As a consequence, healthcare data is processed and managed at the regional level by the corresponding regional ministry or Regional Health Service. An interesting example about where this data is available for the research community can be found in Catalunya, where the Catalan Agency for Quality and Health Evaluation (*Agencia Qualitat I Avaluacio Sanitaries de Catalunya, AQuAS*)[8] through the programme PADRIS allows researchers to access the data generated by the Catalan healthcare system. AQuAS is a Catalan regional government's public agency; its mission is to generate knowledge through data analysis.

4 Education and Training

Just as in healthcare, the education system is decentralised at the Autonomous region level. The regional education authorities ("Consejerias") maintain comprehensive information systems that include data across all levels of the education system. This may include, among others, student admissions and their socioeconomic backgrounds, scholarships, school attendance, grades, etc. (Gortazar & Ferrer, 2023). However, this data has primarily been utilised by education authorities for administrative and monitoring purposes, thereby limiting its availability to the broader research community. While there are some national databases that provide harmonised administrative microdata across regions in the domain of healthcare, similar centralised databases for education do not currently exist. Conversely, both regional and national governments do publish aggregated indicators that can serve to monitor the education system's evolution. One example is the National System of Education Indicators.[9]

However, the utilisation of education microdata for impact evaluation remains very limited. There are instances of certain regions in which education microdata has also been used for evaluation purposes, either internally within education institutions or through collaborations with researchers. One such region is the Basque Country, in which the Basque Institute of Evaluation and Research in Education (*Instituto Vasco de Evaluación e Investigación Educativa*, ISEI-IVEI) has promoted several studies on multilingual education, equity, and learning losses during the COVID-19 pandemic, among others. For instance, Arenas and Gortazar (2024) use data on external assessments in primary and secondary schools to study the learning losses due to school closures during the Covid-19 pandemic. They find learning losses

[8] Access can be granted to researchers from Catalan Universities and research centres. More information about how to access AQuAS data can be found (also in English) at: Analítica de dades—PADRIS. Agència de Qualitat i Avaluació Sanitàries de Catalunya (AQuAS) (gencat.cat).

[9] Sistema estatal de indicadores de la educación: https://www.educacionfpydeportes.gob.es/inee/indicadores/sistema-estatal.html.

equivalent to 13% of the usual learning that would take place in one regular year, a significant smaller magnitude than that found in other countries in the shorter term.[10] Other regions such as Andalusia, Balearic Islands, Canary Islands, or Madrid have occasionally carried out external assessments that have been used for the evaluation of several education policies.

While these examples demonstrate progress in leveraging education microdata for evidence-based policymaking in Spain, they remain somewhat isolated cases that are scattered across different regions and are often not consolidated over time. Moreover, access criteria to the data are generally not transparent and are restricted to a limited number of actors in the education policy evaluation environment. A more harmonised and consistent approach that guarantees transparent access to the data is necessary in order to fully harness the potential of education microdata for impact evaluation.

5 Income, Wealth, and Firm-Level Data

5.1 Income and Wealth

The Institute of Fiscal Studies (IFS), in collaboration with the Tax Authority (AEAT), have been producing datasets for researchers that contain individual-level information extracted from income tax declarations for over three decades. The IFS currently offers access to three different categories of microdatasets with tax declaration information.[11] Access to these databases is available to all researchers free of charge.

- **Samples from Income and Wealth Taxes**: Two random samples that are representative of tax-filers of income and wealth tax declarations, respectively. The period covered by the samples obtained from the income tax declarations currently spans 2000–2021, while that obtained from the wealth tax date back to 2003. Importantly, the IFS provides a sample of individuals who do not have, or did not declare, income tax thanks to the information that is submitted by their employers.
- **Income and Wealth Panel**[12]: Two different datasets containing information from both income and wealth tax declarations of individuals who are followed over a specific period. There are currently two separate panels that contain information from the income-tax declarations, one spanning from 1982 to 1998 and another from 1999 to 2016. The panel with wealth information spans 2012–2016. In

[10] Gortazar and Ferrer (2023) provide an extensive list of evaluations undertaken with regional external assessments data and other administrative education microdata in Spain.

[11] Link to the IFS ant Tax authority with all the available datasets: https://www.ief.es/Investigacion/areaEstadistica.vbhtml and: https://sede.agenciatributaria.gob.es/Sede/estadisticas/muestras-datos-fiscales.html. A brief description of the databases is generally available in English on the website, although the core data documents are provided only in Spanish.

[12] Known in Spanish as *Panel de Renta* and *Panel de Patrimonio*, respectively.

Agrawal et al. (2024), the authors use a linked version of both panels to study the effects of decentralised wealth taxation on mobility and the effectiveness of tax coordination at mitigating tax competition.

– **Household Panel**[13]: The latest dataset made available by the IFS, in collaboration with the Tax Authority and the National Statistics Institute (INE), contains a representative sample of 5.7% of Spanish households for the period 2016–2021.[14] Two main features that are worth highlighting here are: first, the population of reference includes households, instead of single tax-filers, even though the lowest level of information that is available is each single tax filer belonging to the household sampled. The second difference is that it combines information coming from both income and wealth tax declarations. Households are selected through a stratified random sampling procedure in which the strata variables used included region of residence, household type, and income category.[15] The households that were first sampled to appear in the panel in 2016 are then followed over time and new households from the pool of excluded ones are resampled in each wave, in order to preserve the panel's representativeness. The files related to the income tax declaration contain highly detailed information about all of the different sources of income and tax deductions declared by each tax filer in the household sampled. Wealth tax declarations, which only a small share of households must fill in, provide information about the properties and financial assets held by each household.

Groups of researchers have recently gained access to a part of the financial information contained in the records of some commercial banks. Access to personal accounts allows researchers to observe individual-level information about the different sources of income as well as the total amounts held. The main benefit of this data over that of income tax records is that its information is offered at a higher frequency and with a lower delay.[16] Different research areas benefit from access to this type of data; from research devoted to understanding the determinants of income and wealth inequality to that devoted to measuring the effects of social benefits on employment status. A good example of the use of this data for research can be found in Aspachs et al. (2022) which relies on anonymised data from bank records to track inequality.

Access to this data signifies a great improvement for research; however, we believe that there is room for improvement in several dimensions, given its current state. First, the lack of transparency concerning the availability of this privately held data, together with the discretion of their holders in selecting the research projects,

[13] Known as *Panel de Hogares* in Spanish.

[14] Households from the regions of the Basque Country and Navarre are not included in this data as the information held in their regional Tax Authorities is not shared with the national Tax Authority.

[15] If regional-level representation is not required, then the user can instead rely on a subsample with information that is representative of the selected household types and levels of income at the national level.

[16] This data can also provide a better picture of the financial assets held by individuals, since only realised gains are taxed in the year when these assets are liquidated.

might create unequal access to the data by researchers. Second, the data held by one commercial bank might lack the necessary geographic and social representativeness for policy relevance. Third, the cost of setting up a secure IT process by which to provide access to a broad base of researchers might be too high for most commercial banks. This is why creating a representative sample of all of the country's bank accounts should be considered by policy makers as a tool to harness the value contained in this data for research. The task to both coordinate the collection of a representative sample of bank accounts and to make it available to researchers in a secure way could be undertaken by an independent institution or by a branch of the government. Such an effort would make access to this data both more efficient and equitable while also increasing its quality and maintaining the necessary level of security.

5.2 Firm-Level Data

Access to financial and economic information at the firm level has been made possible via the Bank of Spain's data laboratory (BELab).[17] The Bank of Spain offers the possibility for researchers to access several datasets with firm-level information over a considerable period. The researcher might have to access this information in a secure room in the Bank of Spain's premises or remotely, depending on the level of security required by the data sought.

– Microdata on Individual Firms[18]: It contains a panel with all non-financial businesses located in Spain and information on more than 20 characteristics for each of them. The panel contains information from 1995 and it is updated twice a year. The data was originally obtained from the mandatory annual financial statements submitted by each business to the Commercial Registers. There are roughly 900,000 businesses included in the last years of the panel and these account for more than 50 per cent of the Spanish economy's total gross value added and employment. This data has been used in Almunia and Lopez-Rodriguez (2018), for instance, in order to study the effects of size-dependent tax enforcement on firms' tax compliance.

The benefits of garnering firm-level information can be significantly enhanced by linking their data files with the worker-level Social Security microdata records. Through such a connection, more, higher quality research could be performed when assessing the impact of public policies on different domains such as on the labour market, taxation, and competition policies on firm productivity.

[17] More information about the available datasets and access applications can be found at the BELab website: https://www.bde.es/wbe/es/para-ciudadano/servicios/belab/.

[18] Microdatos de empresas individuales (CBI) in Spanish.

6 A Promising Way Forward? New Initiatives to Merge Datasets

Despite Spain having traditionally lagged behind other Member States in guaranteeing access to microdata for research, recently undertaken initiatives are expected to improve this situation. The Spanish Tax Authority (AEAT), the Social Security Administration, the Bank of Spain, and the State Public Employment Service (SEPE) have all recently signed an agreement that enables joint access to their databases for scientific and public interest research. This agreement is expected to allow for the linkage of microdata from different relevant data sources with information on many domains for both individuals as well as for firms. This would enable the effectiveness of relevant policies to be evaluated in a comprehensive and multidimensional matter.

Notably, with the aim of expanding the scope of available data, the agreement is open to other public institutions that may want to make their administrative data available for the research community. Data is made available to the research community; to access it, the organisation with which the researcher is affiliated needs to register as a research institution. A research proposal must be submitted in order to achieve this. Its validity will be assessed by a Committee within 8 to 10 weeks. After that, access to microdata would be guaranteed in a safe centre of one of the institutions involved in the project.[19]

Another important upcoming initiative includes the introduction of national standardised tests for all students ages 10 and 14. This data would help to monitor the evolution of students' competencies, as well as the evaluation of education policies, particularly if access to microdata is facilitated to researchers.[20]

7 Conclusions

Administrative data sources in Spain are dispersed throughout different institutions and at various levels of governmental jurisdictions. These databases contain a wealth of information that, if effectively linked, could unlock significant potential for evidence-based policymaking and enhanced population well-being. While certain institutions have made substantial efforts to enable access to their data records, there is a clear need for a national strategy to both link and facilitate access to administrative records. Indeed, the recent initiative to facilitate access and to link the administrative records held within the four institutions mentioned above should serve as a major starting point. A national strategy is still needed, however, in order to assist or to simply nudge the remaining institutions that hold administrative records and that lack the necessary culture, human capital, or IT infrastructure in order to facilitate their access for research purposes.

[19] For more information on this project, see: https://www.es-datalab.es/.

[20] https://elpais.com/educacion/2024-02-25/el-nuevo-informe-pisa-espanol-examinara-en-abril-de-lengua-matematicas-y-ciencias-a-casi-un-millon-de-estudiantes.html

The government system's decentralised nature may indeed create challenges when attempting to both link and harmonise administrative data for research and policy evaluation purposes. However, this should not pose a barrier to maintaining and providing comparable microdatabases across the country. A case in point is the Basic Minimum Set of Data, where harmonised hospitalisation data is collected by hospitals and is later shared through the National Statistics Institute. This shows that such harmonisation and reporting processes are possible and useful in providing insights into policies that are managed at the sub-national level. This could serve as an example for the necessity to develop similar databases in other decentralised sectors, such as education, long-term care, and for other social policies.

Indeed, decentralisation also provides other opportunities for policy evaluation and evidence-based policymaking, given that it allows each autonomous region to implement distinct public policies. As a result, the use of administrative data for policy evaluation at the regional level can offer valuable insights into the efficacy of innovative policies. Understanding both the successes and shortcomings of these regional initiatives can provide valuable lessons for all regions, ultimately informing the development and implementation of more effective public policies throughout the country.

While each country has its own unique characteristics, examining the experiences of other European Member States can provide valuable insights into the type of infrastructure needed for administrative data sharing in Spain. In general, the establishment of a specialised, third authority or the utilisation of the national Institute of Statistics in the linking, preparing, and sharing of administrative data can both significantly enhance efficiency and optimise resources. Centralising this responsibility eliminates the need for each institution to develop its own protocols and infrastructure to access and share data, thereby streamlining the process and avoiding duplications. This approach may increase the efficiency of resources when compared to a more decentralised system, ultimately facilitating smoother administrative data sharing and utilisation for evidence-based policymaking.

References

Agrawal, D. R., Foremny, D., & Martinez-Toledano, C. (2024). Wealth tax mobility and tax coordination. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4793904

AIREF. (2020). *Opinión para una estrategia de acceso a datos administrativos* (1/20). Autoridad Independiente de Responsabilidad Fiscal.

Almunia, M., & Lopez-Rodriguez, D. (2018). Under the radar: The effects of monitoring firms on tax compliance. *American Economic Journal: Economic Policy, 10*(1), 1–38. https://doi.org/10.1257/pol.20160229

Arenas, A., & Gortazar, L. (2024). Learning loss one year after school closures: Evidence from the Basque Country. *Series*. https://doi.org/10.1007/s13209-024-00296-4

Aspachs, O., Durante, R., Graziano, A., Mestres, J., Montalvo, J. G., & Reynal-Querol, M. (2022). Real-time inequality and the welfare state in motion: Evidence from COVID-19 in Spain. *Economic Policy, 37*(109), 165–199. https://doi.org/10.1093/epolic/eiac008

Borra, C., Costa-Ramón, A., González, L., & Sevilla-Sanz, A. (2021). The causal effect of an income shock on children's human capital. *Barcelona GSE Working Paper Series*, *1272*. https://www.zora.uzh.ch/id/eprint/205603/

De Quinto, A., Hospido, L., & Sanz, C. (2021). The child penalty: Evidence from Spain. *Series, 12*(4), 585–606. https://doi.org/10.1007/s13209-021-00241-9

Farré, L., & González, L. (2019). Does paternity leave reduce fertility? *Journal of Public Economics, 172*, 52–66. https://doi.org/10.1016/j.jpubeco.2018.12.002

Garcia-Louzao, J., Hospido, L., & Ruggieri, A. (2023). Dual returns to experience. *Labour Economics, 80*, Article 102290. https://doi.org/10.1016/j.labeco.2022.102290

Garcia-Pérez, J. I., Serrano-Alarcón, M., & Vall-Castelló, J. (2024). Long-term unemployment subsidies and middle-aged disadvantaged workers' health. *Journal of Population Economics, 37*(1), 14. https://doi.org/10.1007/s00148-024-01000-3

González, L., & Rodríguez-González, A. (2021). Inequality in mortality in Spain. *Fiscal Studies, 42*(1), 103–121.

González, L., & Trommlerová, S. K. (2023). Cash transfers and fertility: How the introduction and cancellation of a child benefit affected births and abortions. *Journal of Human Resources, 58*(3), 783–818. https://doi.org/10.3368/jhr.59.1.0220-10725R2

Gortazar, L., & Ferrer, Á. (2023). *Ocho propuestas para que el sistema educativo no se quede atrás en la revolución de los datos*. https://doi.org/10.56269/20230510/LC

Güell, M., & Lafuente, C. (2022). Revisiting the determinants of unemployment duration: Variance decomposition à la ABS in Spain. *Labour Economics, 78*, Article 102233. https://doi.org/10.1016/j.labeco.2022.102233

Holub, F., Hospido, L., & Wagner, U. J. (2021). Urban air pollution and sick leaves: Evidence from social security data. *SSRN Electronic Journal*, *72*.

Juanmarti Mestres, A., López Casasnovas, G., & Vall Castelló, J. (2021). The deadly effects of losing health insurance. *European Economic Review, 131*, Article 103608. https://doi.org/10.1016/j.euroecorev.2020.103608

Marie, O., & Vall Castelló, J. (2023). Sick leave cuts and (unhealthy) returns to work. *Journal of Labor Economics, 41*(4), 923–956. https://doi.org/10.1086/720629

Roca, J. D. L., & Puga, D. (2017). Learning by working in big cities. *The Review of Economic Studies, 84*(1), 106–142. https://doi.org/10.1093/restud/rdw031

Josep Amer-Mestre is a Quantitative Policy Analyst at the Competence Centre on Microeconomic Evaluation (CC-ME) of the Joint Research Centre of the European Commission. His research focuses on evaluating public policies in the fields of education and the labour market. Josep holds a Ph.D. in Economics from the European University Institute (EUI).

Manuel Serrano-Alarcón is a Quantitative Policy Analyst at the Competence Centre on Microeconomic Evaluation (CC-ME) of the Joint Research Centre of the European Commission. His research interests revolve around the counterfactual impact evaluation of public policies, particularly in the areas of social and health policies. He holds a Ph.D. in Public Health/Health Economics from NOVA University of Lisbon, an M.Sc. in Economics from the University of Warwick and Lund University, and a B.Sc. in Economics from the University of Alicante.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 28
Accessing Administrative Data for Research: Lessons from Sweden



Stefano Lombardi

Abstract This chapter focuses on the most common and flexible way by which to access Swedish microdata in order to conduct research in economics and the other social sciences, which is via Statistics Sweden. The system appears to be well-developed and equal for two reasons. First, data access is managed by one authority and the rules determining how data is accessed remotely are clear. Second, when applying to access microdata, an assessment is made concerning the proposal's legal feasibility, without entering into specific details of whether and how the research proposed is scientifically promising. Access to administrative data registers at Statistics Sweden is given provided that an agreement is stricken between Statistics Sweden and a Swedish institution. Therefore, foreign-based researchers who are interested in using Swedish microdata need to undertake a collaboration with Sweden-based researchers and must have a research agreement with a Swedish institution.

1 Overview

This chapter has two aims. First, it provides information about how to access Swedish administrative data for research purposes. The focus of the discussion is on those data sources that are generally used to do research in economics and other social sciences. However, many of the messages presented in this chapter, as well as the examination of the underlying regulation, can be applied more generally. Second, the chapter outlines the data protection regulations that are currently in place and evaluates the system's fairness in terms of granting researchers access to the data. The chapter is tailored to the needs of both researchers and policymakers. Researchers will gain an understanding of the types of register-based data that are available in Sweden, while policymakers in Sweden and beyond will receive a practical overview of how these regulations safeguard the data, while balancing the need to provide public access to administrative data for the greater good.

S. Lombardi (✉)
VATT Institute for Economic Research, Helsinki, Finland
e-mail: stefano.lombardi@vatt.fi

© The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_28

Administrative registers are unit-level data (individuals, households or firms) in which units are followed over time (most often at the yearly level, but possibly up to the daily level). Sweden, like many other Nordic countries, is at the forefront of providing access to microdata for research purposes. One important reason for this is that Sweden has a long tradition of digitisation and data infrastructure management. Since administrative data are data that were originally collected for reasons other than for scientific research and statistical surveys, having a well-developed pre-existing data infrastructure has certainly helped to both gather and organise high-quality data sources in recent decades. A political willingness to develop a system in which certain costly policies (such as active labour market policies for the unemployed) are properly evaluated also explains why Sweden, much like other Nordic countries, has a highly developed system of data access.

Nordic countries share important similarities, but also some differences among each other, when it comes to granting access to microdata. Therefore, the reader might also be interested in comparing the data access conditions for Sweden with those for Denmark (Chap. 8) and Finland (Chap. 10), which are also covered in this book.[1] As for the Finnish case, this chapter provides details about some existing data sources, but does not attempt to provide a comprehensive list of administrative data. Instead, the goal of the discussion is to provide details about the most common and flexible way through which a researcher can get access to Swedish microdata, which is via Statistics Sweden (*Statistiska centralbyrån*, henceforth SCB), the Swedish National Statistical Authority.[2] SCB's microdata can be used for scientific research and can be accessed remotely via secure two-factor authentication. The next section broadly describes those data sources that can be accessed at SCB.

The remainder of the chapter is organised as follows. Section 2 describes different administrative data sources. Section 3 presents information about how to access administrative data at Statistics Sweden. Section 4 describes the data protection rules that are currently in place. Section 5 presents an assessment of the equality of data access and Sect. 6 concludes with a discussion of data access from a policy perspective.

2 Data Sources

All microdata accessed from SCB's servers can be combined by using protected and unique unit identifiers (pseudoidentifiers). SCB creates project-specific pseudoidentifiers that are valid only within a given project and that are destroyed when the project is completed. Pseudoidentifiers allow researchers to follow individuals over time and to link their information across registers. As is the case for other Nordic

[1] See also Atella et al. (2025), which provides information on access to health registers in the Finnish context.

[2] SCB is a Swedish state administrative authority that falls under the Ministry of Finance and is responsible for official statistics and other state statistics.

countries, workers can be linked to firms and establishments, co-workers to their peers and individuals to spouses and other family members across generations. In addition, detailed firm registers with yearly information from balance sheets and personnel data can be linked to the employers over time.

This information can be used to analyse several aspects of people's socioeconomic trajectories throughout their lifetime. For instance, worker- and employer-level characteristics and daily unemployment and programme spell data can be combined in order to analyse the effectiveness of labour market policies (Lombardi et al., 2018, 2024; van den Berg & Vikström, 2022) and to study unemployment insurance systems (Lombardi, 2019; van den Berg & Vikström, 2014). The matched employer–employee nature of the data can also be used to study the consequences of job loss (e.g., Bertheau et al., 2023), wage inequality (Åslund et al., 2023) and the functioning of labour markets for different networks and groups of workers (Bratu et al., 2020; Kramarz & Nordström Skans, 2014). Detailed education information and people's demographics further allow for education institutions to be studied (e.g., Fredriksson & Öckert, 2014; Fredriksson et al., 2013). Moreover, linking detailed individual-level data about earnings and other socioeconomic characteristics across generations allows researchers to assess the extent of intergenerational income mobility (Adermon et al., 2021; Bratu & Bolotnyy, 2023).

SCB provides access to a wide range of administrative registers that are ready to be used (*ready-made data*). It is also possible to require data to be pre-processed and linked to other registers, based on the user's needs (*tailored data*). Moreover, and under specific conditions, data delivered from external organisations (including population-level microdata) or self-collected data can also be linked to SCB's registers (*external data*). The possibility of pseudonymising and linking external data to the other databases used in a project makes the overall system very flexible.

Ready-made registers are register databases that have been cleaned and made available for research and statistical purposes via SCB's servers. Many research questions in economics and social sciences can be answered by using this source of information and without the need to rely on externally linked data. Ready-made registers include very detailed, population-level, longitudinal information on both individuals and employers.[3] The information available is divided into the following categories on SCB's website (Swedish version)[4]:

- Labour Market:
  - Link between company-level information and shareholders (entrepreneur, owner);
  - Information on sickness, parental and unemployment insurance at the individual level;

[3] Several registers, such as old census data, have not been fully digitised yet. Depending on researchers' time and resources, digitalisation efforts can allow for the available data to be augmented further.

[4] In the interest of space, the focus here is on administrative registers that cover the entire population or at least large representative samples thereof. Additional survey-based resources are available from SCB.

  - Detailed set of individual-level demographics and income characteristics measured at the end of the year and at the employment spell level;
  - Link between a person and their family, workplace and firm;
  - Wage level and structure in the private and public sector.

- Demography:
  - Multigenerational links of people born after 1931 and registered in Sweden since 1961 (both with biological and with adoptive parents);
  - Demographics register (since 1968) with information on gender, age, marital status, citizenship, date of birth and death and immigration and emigration episodes, among others.

- Property:
  - Microdata about all properties, buildings, addresses and apartments in Sweden;
  - Property price register on sales and other transfers of properties, for all types of properties

- Business:
  - Detailed firm and workplace balance sheet information;
  - Energy use, company expenditures on IT, foreign trade in goods, R&D register.

- Household finances and expenses
- Income and taxation
  - Register on financial assets and liabilities;
  - Income and taxation register;
  - Register of capital gains and losses.

- Education
  - Education enrolment;
  - Final grades from upper secondary school;
  - Teacher-level information;
  - Doctoral students' information, staff at universities and college, university students' grades and other information, applicants and admitted to university;
  - Vocational school's register for applicants, students and graduates.

The databases are generally available at the yearly level from the beginning of the 1990s. Researchers can also access census registers that were collected every five years between 1960 and 1990 (when the yearly population registers start being made available).

Tailored data consists of data that is produced by SCB's statisticians based on researchers' needs. It is possible to request data produced in a specific format or for certain groups of units upon paying a processing fee. This information can be linked to existing ready-made registers and accessed via SCB's servers. Moreover, it is possible to submit a pre-defined list of individuals to which SCB then links the register data. SCB can also draw a control population to match to a case population that has been defined by the user.

External data are data sources that are linked to SCB's registers. If an external organisation needs to deliver data for an SCB project, then the user is responsible for placing separate orders for both the data and for managing the contact and conveying the necessary information between SCB and the external data holder. The two sources must be connected in order for SCB to be able to provide access to register data with data from other register holders. This can be done in two ways: (i) the external data holder delivers data to SCB, and SCB replaces the identity with a serial number; (ii) SCB sends a code key (a link between identity and serial number) to the record holder. SCB can only do this if the record holder is subject to statistical confidentiality. The secure submission of data occurs via an SFTP account, which is set up by SCB. In both cases, the researcher and the organisation who provided the external data sources need to have a formal legal agreement about the data's usage. This step can be time-consuming, depending on the data source, especially in the event that ethical approval for the use of the data is required (see next section).

3 The Application Process for Data Access

3.1 Overview of SCB's Application System

In order to access SCB's microdata, the applicant needs to electronically submit a research plan that clarifies why the data requested is necessary for the purpose of answering the research questions of interest. This concerns both the variable groups required and the use of population-wide registers. The data must be appropriate, relevant and not too extensive for the purpose. The legal assessment of whether a user licence can be issued for the specific application is not based upon the justification of the scientific novelty of the research question in itself, but rather on the data protection legislation that is in place.

Statistics Sweden's registers contain, among other things, information about political opinions, information that can be linked to membership in a trade union, income information linked to health and information that can indicate a functional impairment. If the application includes these, or other personal data considered sensitive according to the GDPR, then the applicant must have a project approval from the Ethics Review Authority in order for sensitive personal data to be processed.[5]

[5] Sensitive data include ethnic background, political opinions (political party affiliation, for example), religious or philosophical beliefs, membership in a trade union, health, a person's sex

Applications are processed on a first come, first served basis. When SCB receives an application, then the user receives an estimate of the current queue time and will be assigned a case manager as quickly as possible. It normally takes 1–6 months from the time a person is assigned to a case manager to the time when data can be accessed.[6] Information about the approximate waiting time can be found on SCB's website. If sensitive data needs to be processed, then the user does not need to have approval from the Ethics Review Authority to place an order and to join the queue, but the ethical approval needs to be submitted eventually. However, the application can be assigned to a case manager at SCB only when the Ethics Review Authority decision is made available.

Several factors affect the price and how long it takes to process an application. In particular:

- How clearly the application describes the population and requested register data, and whether there is support for processing them;
- How quickly the person making the order answers the administrator's questions about the order;
- The complexity of the population or sample to build;
- The number of records requested;
- How many different identities need to be replaced with pseudoidentifiers;
- Whether the user requires external data to be linked to SCB registers and the size and complexity of the external registers;
- Whether SCB is dependent upon other register holders to move forward (for instance, when the external data holder needs to create the target group of interest).

After having approved the confidentiality review, the user will receive a fixed price quote and SCB's commitment in terms of time of delivery. As of 2024, SCB reports that most of the applications fall within the cost range of 35,000–65,000 SEK excluding VAT (3,000–5,500 EUR). In general, the size of the population and the number of requested register data affect the price marginally if the data can be retrieved from only one register. There are different ways to access data; the primary way is remotely and via the MONA secure connection system.

3.2 Conditions to Access Statistics Sweden's Microdata

There are two groups of authorised entities that can purchase microdata at SCB:

life or sexual orientation, genetic information and biometric data that could be used to uniquely identify a person.

[6] From the researcher's perspective, the time needed to open an SCB project can be considerably longer in cases where data permits from external data holders and/or an ethical approval for processing personal information are needed. This is the case, for instance, when the researcher is interested in linking external health register data.

- Swedish universities, colleges, authorities, regions and private organisations that request microdata for a specific delimited project with a research purpose;
- Authorities responsible for statistics and other public organisations that request microdata for limited projects for statistical analyses and reports with a statistical purpose.

SCB does not release microdata with purposes other than carrying out research and statistical analysis. For example, SCB does not disclose data to either analyse a customer group for commercial purposes or in order to investigate or audit individuals or companies.

In practice, it is a Principal Investigator (PI) that sends the electronic application to SCB. However, the data recipient is always a Swedish institution, which signs a personal data processing agreement with SCB (data accessed remotely is also considered to have been transferred). Whoever acts as a PI needs to have an affiliation or other type of agreement with a recognised Swedish institution.

Who can be a data processor? Researchers affiliated with a Swedish institution can access microdata via their affiliation. For researchers working in non-Swedish institutions, it is possible to access SCB's data if the foreign-based researcher works in a research team for which the PI is either employed or affiliated with a Swedish university. In this case, data is accessed from a foreign institution in the following way. First, as mentioned previously, the data must be ordered and guaranteed by the Swedish research institute. The project must also get secrecy clearance from SCB before the data is made available. Next, after SCB gives the secrecy clearance to the Swedish university, the Swedish university must give the secrecy clearance to the foreign organisation from which the foreign-based researcher works. It is then up to the Swedish university to decide if an affiliation is necessary or if the data can be released to the foreign university without affiliation requirements.

The assessment of how much data it is possible to apply for with a given user license is always made on a case-by-case basis. What matters for the final decision are data protection considerations on whether the final set of data, which is linked together, is in line with data protection laws (see Sect. 4) and not the data source used (ready-made, tailored and external). Therefore, in principle, it is possible to link together SCB's registers with external administrative data, provided that the data linkage is justified by the research question and is in line with the law.

From the perspective of linking external data sources to SCB's registers, understandably, the institution that provides the external data always has a certain degree of discretion in accepting delivering the data for research use. At the same time, there is a long tradition of using administrative data to do policy evaluation and research in general in Sweden. Therefore, in many cases, it is possible to obtain very detailed external data sources, often at no cost.

3.3 Alternative Ways to Access Data

Since SCB has data delivery agreements with institutions, it is possible for Swedish research centres to have projects that their affiliates can join, with data being shipped to the research centre's servers and accessed from there (including remotely). In these cases, all usual data protection and confidentiality rules must be met. However, from the perspective of a researcher, the time and costs to open/access the data are considerably lower compared to what is required in a standard SCB project.

While the conditions through which data can be accessed in these cases vary on a case-by-case basis, from the perspective of a foreign-based researcher interested in accessing Swedish microdata, it might be sufficient to be part of a research team with a PI that is affiliated with the Swedish research institute that has the agreement with SCB. In this sense, the procedure to get access to the data is no different from a standard SCB project. Of course, as usual, both the researcher and the non-Swedish institution where the foreign-based researcher works from need to have security clearance in place.

4 Data Protection

This section discusses how scientific research that uses microdata, which is a task of public interest, can be performed in the Swedish context and in accordance with data protection rules.

4.1 SCB Confidentiality Assessment and the Personal Data Processing Agreement

In general, information about individuals and companies contained in SCB's registers is protected by absolute confidentiality, according to Chap. 24, Sect. 8 of the Publicity and Secrecy Act (2009: 400). However, SCB can make exceptions to this rule and release data for research or statistical projects. The disclosure of information can only take place for a specific, defined project and, as a rule, identities such as social security numbers and organisation numbers are always replaced with serial numbers (pseudoidentifiers).

Before SCB releases the data, it must perform a so-called confidentiality test to make sure that data protection is in place. In practice, prior to the confidentiality review, SCB administrator assigned to the case and the applicant have an ongoing conversation in order to determine what is being requested and that there is support for the information to be released. When the administrator and the applicant agree on what the assignment includes, the administrator summarises the assignment in a

confidentiality review and sends it on for decision. A formal decision on disclosure is made by Statistics Sweden's unit manager for assignments and external service.

The decision on whether the information can be released is based on several factors. SCB investigates whether there is support for breaking the confidentiality and for releasing the requested information. A prerequisite is that the information being disclosed is covered by an equally strong confidentiality protection from the recipient. Organisations that SCB has not previously disclosed information to must first describe their technical and organisational safeguards before SCB can decide on whether the information can be disclosed, subject to confidentiality. Information covered by confidentiality must never be disclosed to private individuals, not even with a confidentiality clause. Moreover, SCB examines whether there is a risk that, after disclosure, the data may be processed in violation of (i) data protection regulation (GDPR); (ii) the law (2018: 218) with supplementary provisions to the EU's data protection regulation and (iii) Section 6 of the Ethics Review Act. This test includes an assessment of whether the processing meets the data protection regulation's requirements for personal data minimisation.

SCB projects are temporary and can be extended, upon request, for specific reasons. All data processors agree to abide by data protection rules throughout the duration of the project and confidentiality rules must be followed, even after the project's completion. Users found to have infringed the rules are subject to termination of data access, even for future applications (for individual researchers and possibly even for their organisation).

When SCB releases the microdata, the recipient (the organisation to which the data was released) becomes responsible for that data, irrespective of whether the data was released via MONA or in some other way. This happens only after a personal data processing agreement regarding the personal data is signed between SCB and the recipient. If the recipient is an authority, then the recipient decides which employees within the organisation may have access to the data. People outside the authority can access data remotely only if they are linked to the authority, as though they were employees. If the information is to be released to another organisation, then the authority must carry out its own confidentiality assessment. In cases where the microdata is released to another organisation, MONA remote access can only be used in special cases. In practice, this can limit data access for people working in non-Swedish institutions, although it does not prevent them from doing so. Private organisations can also be data recipients, according to specific rules.

4.2 Remote Access and Exporting Output Data

The requested information is provided in the form of microdata once the confidentiality assessment has been approved. The recipient is responsible for disclosed information. It is up to the recipient to protect the data from unauthorised access and to prevent dissemination. SCB offers the MONA platform in order to make this

possible and through the microdata can be processed together with data from other sources in a restricted and secure environment with two-factor login.

SCB only delivers data in MONA without personally identifiable information, in accordance with the GDPR and in order to secure integrity and confidentiality; personal identity numbers are replaced by serial numbers prior to delivery. Moreover, personal data processing is limited to users who have been approved by the recipient's institution. Nevertheless, the released data is personal data and must be treated as such. The data is protected by secrecy and the recipient shall ensure that the information is not disseminated to unauthorised parties and may not be used by other organisations or for other purposes. MONA, SCB's remote access system to administrative registers, is a solution that ensures such processing.

Information in MONA is processed in accordance with the terms of use that stipulate, among other things, places from which MONA may be used. MONA can only be used in the EU, EEA and in third countries that have been approved for transfer by the European Commission. MONA's terms of use and additional documentation are available on SCB's website.

In practice, MONA users work in a Windows environment via a remote connection and the microdata is displayed on the computer screen. Users can only download results to their own computer; the microdata must always stay safely stored in MONA. The system is flexible, given that users have the possibility of transferring their own material to their storage area in MONA, such as aggregate statistics or scripts necessary for the purposes for which SCB released the data. MONA users have access to metadata regarding most of Statistics Sweden's microdata registers in addition to having access to released microdata, software and their own storage area. Most metadata and documentation, including on SCB's website, is in Swedish.

Particular attention must be paid to the secrecy requirements when it comes to exporting results. According to the terms of use, it is the researcher's responsibility not to download any data that contains microdata from MONA. Microdata refers to data on individual units (e.g., people, companies, workplaces, schools, etc.). Users are allowed to export aggregated data (i.e., cell-level data that contains only few objects), but it is the user's responsibility to verify the information before disclosing it, given that the data is protected by secrecy. SCB offers guidance on what conditions need to be fulfilled when exporting data thorough a report (in Swedish) about disclosure control.[7]

5 Equality of Data Access

The system appears to be quite well-developed and equal in many respects. The system is centred around the national statistical authority and the rules for accessing the data remotely in a secure environment are clear. In addition, the data is released

[7] See here: https://www.scb.se/contentassets/0cd92207266d40eb8829244d51d90b94/handbok-i-statistisk-rojandekontroll.pdf.

for research purposes when applying to access microdata at SCB, but the decision about whether to issue the data permit is not taken by a scientific committee. Instead, the decision about whether to provide access to the microdata is taken by statisticians and lawyers, based on data protection rules, and the application is evaluated only on the basis of its legal feasibility; this is similar to what happens in Finland and in other Nordic countries, for instance. This also applies to projects that require the use of external microdata that is linked to SCB's registers. Overall, the system is extremely flexible in terms of what data sources can be used for research purposes. Data access is also flexible, given that the registers can be easily accessed from any EU and EEA country, or from other countries that have been recognised as secure according to GDPR.

Data access is conditional on fulfilling specific data security requirements from the perspective of a researcher interested in using Swedish administrative registers via SCB. In particular, any data access agreement is always made between SCB and a Swedish institution that must, in turn, have an agreement with individual researchers. While this is not a practical issue for researchers employed at Swedish universities, obtaining data access for people employed at foreign-based institutions is not a given. According to what is reported on their own website, SCB is strict in terms of providing access to researchers employed in foreign-based institutions. However, an increasing number of non-Swedish universities have become part of agreements with SCB and Swedish universities over time, thereby making it easier for foreign-based researchers to work with Swedish data in the context of international collaborations. It is reasonable to expect that this trend will continue into the future.

From the perspective of foreign-based researchers, the requirement to have established connections with Swedish-based researchers might appear limiting. However, collaborating with researchers based in a Swedish institution to work on Swedish microdata would be very strongly advised even if it was not required by law, given the complexity of the population-wide registers and of the related administrative details. Establishing connections with Swedish-based researchers is also crucial for the purposes of understanding the intricate details involved in the administrative data and concerning the standard data-cleaning and preparation procedures that the Swedish research community has developed over the years. Working together with Swedish-based researchers is also going to be necessary to deal with the fact that large parts of SCB's website and the metadata describing the registers are in the Swedish language. In addition, knowing both where to find and how to apply for external data sources, and whether this is needed at all, is something that can only be learned with time. Some of these external data sources might be hard to obtain without having built trust with the given institution. Therefore, it proves useful and often necessary to work with researchers employed at a Swedish institution in such instances.

6 Conclusions

This chapter has focused on the most common and flexible way by which to access Swedish microdata in order to conduct research in either economics or in other social sciences, which is via SCB's projects. Overall, Sweden, like other Nordic countries, is at the forefront of providing administrative data for research purposes and can provide a useful benchmark for other European countries. The chapter concludes with a discussion of the following policy-relevant questions: *Can the Swedish case be an example for other countries interested in providing access to administrative data? If so, what might policy makers do to make data access possible?*

This discussion is inherently tied to the rules governing data protection. As is evident by comparing across the chapters of this book, the way in which the GDPR is interpreted varies (sometimes considerably) across countries. While Sweden and the other Nordic countries' interpretation of the GDPR, for a mixture of historical and societal reasons, can be considered as being on one end of the spectrum across EU Member States, with comprehensive individual-level data from administrative registers being readily accessible, other countries justify the impossibility of releasing and linking administrative data for research purposes based on the interpretation of the GDPR itself. Such differences in the interpretation of the GDPR suggest that, as of today, it is not obvious where to set the bar when balancing the needs for privacy and for data protection against the fact that evidence-based research that uses administrative data is a public good.

The Swedish case shows an example of how the balance between tasks of public interest and data protection can be achieved in agreement with National and EU laws, while also allowing researchers the flexibility to work with extremely detailed, individual-level information about the whole population. Clear prerequisites include both a substantial amount of trust in institutions and an understanding that high-quality microdata can be useful to reaching this goal, if not an outright necessity to evaluate public policies and to keep building public trust in the institutions via accountability. Much like in other Nordic countries, these prerequisites were gradually built over time. Therefore, if policy makers in other countries are interested in taking inspiration from the Swedish experience, then they should consider importing the Swedish model in a stepwise fashion.

Finally, Swedish policy makers should be aware that, even though this chapter describes a well-functioning system, there is an inherent trade-off between the benefit of having a fully centralised system and the substantial lack of competition in data provision. Other Nordic countries, such as Finland, have shown that the latter aspect could result in an increase in prices, which might have implications for the equality of access.

7 Contacts, Useful Links and Relevant Legislation

Microdata page and research catalogue:

https://www.scb.se/en/services/ordering-data-and-statistics/microdata/
https://www.scb.se/vara-tjanster/bestall-data-och-statistik/register/

Prior to ordering data and how to order data:

https://www.scb.se/vara-tjanster/bestall-data-och-statistik/mikrodata/infor-din-bestallning/
https://www.scb.se/vara-tjanster/bestall-data-och-statistik/mikrodata/bestalla-mikrodata/

Ethical review and sensitive personal data:

https://www.scb.se/vara-tjanster/bestall-data-och-statistik/mikrodata/infor-din-bestallning/vad-tittar-scb-efter-i-etikansokan/
https://www.imy.se/privatperson/dataskydd/introduktion-till-gdpr/vad-ar-person uppgifter/kansliga-personuppgifter/
https://etikprovningsmyndigheten.se/en/

Public Access to Information and Secrecy Act:

https://www.slu.se/globalassets/mw/stod-serv/juridik/public-access-to-inform ation-and-secrecy-act-folder.pdf

Confidentiality review for the disclosure of register data:

https://www.scb.se/vara-tjanster/bestall-data-och-statistik/mikrodata/infor-din-bestallning/sekretessprovning/

MONA remote access details, regulations and user guide:

https://www.scb.se/contentassets/36c49273ef8d4db8a0e2e240510a9a0f/handle dning_eng_20231025.pdf
https://www.scb.se/en/services/ordering-data-and-statistics/microdata/mona--sta tistics-swedens-platform-for-access-to-microdata/rules-and-regulations/terms-of-use/
https://www.scb.se/en/services/ordering-data-and-statistics/microdata/mona--sta tistics-swedens-platform-for-access-to-microdata/rules-and-regulations/
https://www.scb.se/contentassets/36c49273ef8d4db8a0e2e240510a9a0f/handle dning_eng_20231025.pdf

Exporting aggregated data and disclosure control:

https://www.scb.se/hitta-statistik/artiklar/2017/Rojandekontroll--viktig-for-statis tikens-kvalitet/
https://www.scb.se/contentassets/0cd92207266d40eb8829244d51d90b94/han dbok-i-statistisk-rojandekontroll.pdf

SCB contact for questions concerning research data: mikrodata@scb.se.

Acknowledgements I would like to thank Cristina Bratu (VATT Institute for Economic Research) and the research services at Statistics Sweden for useful comments and discussions.

References

Adermon, A., Lindahl, M., & Palme, M. (2021). Dynastic human capital, inequality, and intergenerational mobility. *American Economic Review, 111*(5), 1523–1548.

Atella, V., Ganna, A., & Stefano Lombardi (2025, forthcoming). Optimizing population health through strategic use of health data. *Economia Italiana.*

Åslund, O., Bratu, C., Lombardi, S., & Thoresson, A. (2023). *Firm productivity and immigrant-native earnings disparities* (Vol. 14960). IZA Discussion Paper No. 14960.

Bratu, C., & Bolotnyy, V. (2023). Immigrant intergenerational mobility: A focus on childhood environment. *European Economic Review*, *151*(C).

Bratu, C., Dahlberg, M., Engdahl, M., & Nikolka, T. (2020). Spillover effects of stricter immigration policies. *Journal of Public Economics, 190*, Article 104239.

Bertheau, A., Acabbi, E. M., Barceló, C., Gulyas, A., Lombardi, S., & Saggio, R. (2023). The unequal consequences of job loss across countries. *American Economic Review: Insights, 5*(3), 393–408.

Fredriksson, P., & Öckert, B. (2014). Life-cycle effects of age at school start. *The Economic Journal, 124*(579), 977–1004.

Fredriksson, P., Öckert, B., & Oosterbeek, H. (2013). Long-term effects of class size. *The Quarterly Journal of Economics, 128*(1), 249–285.

Kramarz, F., & Skans, O. N. (2014). When strong ties are strong: Networks and youth labour market entry. *Review of Economic Studies, 81*(3), 1164–1200.

Lombardi, S. (2019). *Threat effects of monitoring and unemployment insurance sanctions: evidence from two reforms*. No. 2019: 22. Working Paper.

Lombardi, S., Skans, O. N., & Vikström, J. (2018). Targeted wage subsidies and firm performance. *Labour Economics, 53,* 33–45.

Lombardi, S., van den Berg, G. J., & Vikström, J. (2024). Empirical Monte Carlo evidence on estimation of timing-of-events models. *Econometric Reviews*, *44*(1), 90–118.

Van den Berg, G. J., & Vikström, J. (2014). Monitoring job offer decisions, punishments, exit to work, and job quality. *The Scandinavian Journal of Economics, 116*(2), 284–334.

Van den Berg, G. J., & Vikström, J. (2022). Long-run effects of dynamically assigned treatments: A new methodology and an evaluation of training effects on earnings. *Econometrica, 90*(3), 1337–1354.

Stefano Lombardi is a senior researcher at the VATT Institute for Economic Research (Helsinki). He received his Ph.D. in economics at Uppsala University in 2019 while being affiliated with IFAU (Institute for Labor Market and Education Policy Evaluation). Previously, Stefano had earned a master's degree in economics from Bocconi University. He is affiliated with IFAU, IZA (Institute of Labor Economics) and UCLS (Uppsala Center for Labor Studies). Stefano's research interests are in applied labour and health economics, with particular focus on the evaluation of labour market policies, the study of the consequences of job loss and the analysis of wage inequality. For his research, Stefano uses population-wide administrative registers from both Sweden and Finland. His research has been published in prestigious journals, such as the

American Economic Review: Insights, *Nature Human Behavior* and *Proceedings of the National Academy of Sciences*.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part II
Horizontal Perspectives

Chapter 29
Cross-Country Networks, Administrative Data and Policy Learning



Josep Amer-Mestre, Sara Flisi, and Paolo Paruolo

Abstract This chapter provides an overview of structured cross-border interactions in the area of data-driven impact evaluations. Interest in these interactions stems from results in the literature that suggest that human and human-data interactions are driving factors for the development of data-driven policy learning. The chapter discusses policy instruments that could be used to support such interactions, may improve coordination and increase public value.

1 Introduction

Exposure to research conducted using administrative data produces positive direct and indirect effects on researchers' productivity, see Nagaraj and Tranchero (2023). The mechanisms at play appear to include *direct* effects, such as the increase in research papers produced thanks to access to registry data, as well as *indirect* effects, such as exposing researchers to evidence inferred from administrative data.

Citizens are attentive to "what works", with new evidence showing that public support for initially unpopular policies can increase once their effects have been demonstrated post-implementation, see, e.g., Carattini et al. (2024). Causal impact evaluation is the cornerstone of evidence-informed policymaking, see EC (2021); it helps to "ensure that decisions are rooted in trustworthy evidence and deliver desired outcomes", and it "promote[s] public accountability, transparency, and contribute[s] to citizens' trust in government", see OECD (2025).

The views expressed in this paper are those of the authors and do not necessarily reflect the opinion of the institutions of affiliation.

J. Amer-Mestre (✉) · S. Flisi · P. Paruolo
European Commission Joint Research Centre, Ispra, Italy
e-mail: josep.amer-mestre@ec.europa.eu

S. Flisi
e-mail: sara.flisi@ec.europa.eu

P. Paruolo
e-mail: paolo.paruolo@ec.europa.eu

© The European Union 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_29

Nevertheless, public administrations face significant challenges in developing robust and reliable monitoring and evaluation systems in practice. They also face reluctance in using the results and evidence of monitoring and evaluations to support future decision-making processes, including budgetary decisions. While some of these challenges may not be resolvable outside the political sphere, this chapter considers the following questions: "can cross-country cooperation improve the quality and use of evidence?", and, in the affirmative case, "what can public policy do to support it?"

To discuss these issues, this chapter starts by listing available cross-country networks, those related both to access to data and to the use of results. It focuses on networks and initiatives that involve both data and people across national borders, including in non-EU countries.

It proceeds by arguing that improvements in the interaction among researchers, practitioners and policymakers on data and results can feed a virtuous circle, thereby enhancing further use of administrative data for causal impact evaluation as well as the spread of the most recent methodological developments. This would lead to better outcomes, both for the research community and for the general public, who would be in a position to learn what works and for whom.

Several policy instruments could be used to foster cross-country and international collaborations, including dedicated funding and a wider circulation of data access opportunities and discussion of results.

The rest of this chapter is organised as follows: Sect. 2 overviews cross-country networks focused on data access and sharing; Sect. 3 reports experiences of cross-country data linkages; Sect. 4 describes fora of exchange involving people; Sect. 5 discusses potential support measures and Sect. 6 concludes.

2 Data Access Networks

This section describes some of the cross-country networks that are currently active in facilitating access to administrative data for research purposes across multiple countries.

Basic steps of working with microdata are first recalled here. At present, researchers seeking to access administrative microdata are usually required to follow a multi-step process, as illustrated in Chaps. 2–28. First, the prospective data users need to identify the specific datasets relevant to their research and the corresponding data-holding institution responsible in each country in the study. Next, researchers need to submit a data access request, usually containing a research proposal with the study's objectives, methodology and data requirements to each of those institutions. If approved, the researchers will then gain access to the data, usually in a secure environment, often through remote data access facilities. The next step involves preparing and harmonising data and performing the analysis.

In the best-case scenario, the data institutions have an established procedure for accessing microdata; in other cases, this procedure may be more ad hoc or informal.

The lack of uniformity in data access in different countries usually makes cross-country studies more difficult and resource-intensive than projects concerning a single country.

Several networks have been created in the last decade to facilitate data access across national borders for research and policy design. Some of them are listed in what follows.

The International Data Access Network (IDAN),[1] founded in 2018, is a collaboration between five Research Data Centres in four different countries. Its current partners are the Secure Access Data Centre (CASD) in France, the Leibniz Institute for the Social Sciences (GESIS) and the Institute for Employment Research (IAB) in Germany, Statistics Netherlands (CBS) in The Netherlands and the UK Data Service (UKDS) in the United Kingdom.

IDAN's mission is to facilitate the research use of secured access data, particularly those coming from administrative records, between its partners via the reciprocal provision of Safe Room Remote Desktop Access. The network promotes remote access to secure data to researchers by establishing a memorandum of collaboration and by signing a bilateral or multilateral memorandum of understanding between its partners.

The network facilitates access to data from multiple countries by comparing and providing harmonised information on each partner's frameworks for secured data access, such as accreditation procedures, requirements for security rules and sanctions in case of confidentiality breaches. All partners have agreed on common security rules so that the IDAN safe rooms can host all partners' access points, thereby allowing the researchers to work with data from several partners from one physical space and respecting all partners' requirements for security. While this simplifies data access across countries through a single *safe access room*, it neither offers the possibility to automatically link data from different countries, nor does it homogenise cross-country data; the single national databases are also not directly comparable.

Another network is the Competitiveness Research Network (CompNet)[2]; it was originally founded by the European System of Central Banks in 2012 in order to foster the debate on competitiveness issues among policy institutions and researchers. CompNet offers researchers micro-founded sets of data covering productivity indicators for some 22 European countries. These data, known as the Vintage CompNet dataset, feature indicators computed by their respective National Data Providers using firm-level data, which are then cross-aggregated and harmonised to allow for cross-country comparability. The different indicators are constructed using information on firms' competitiveness, finance, labour, productivity and trade.

Several summary statistics are provided for each indicator in the dataset, such as selected percentiles, the mean and standard deviation. Currently in its 9th edition, the dataset contains information from 1999 to 2021, although the specific timespan covered differs among countries and it is updated with each new edition. Different

[1] https://idan.network. Last accessed: 1 Feb 2026.

[2] https://www.comp-net.org. Last accessed: 1 Feb 2026.

aggregation levels are also offered such as country, NUTS2, macro-sector and 2-digits industry level.[3] A data request can be submitted by filling out a dedicated form to the CompNet team via the Halle Institute for Economic Research (IWH) Research Data Centre.

CompNet offers a harmonised and reliable source of information on competitiveness and productivity best suited for cross-country comparability. The network is committed to ongoing improvements in both the quality and coverage of its information. Its data aggregation and harmonisation processes are transparent, thereby ensuring that users clearly understand how the information is aggregated and harmonised.

CompNet is also working towards allowing researchers to customise their queries directly on firm-level data instead of on aggregated indicators. This new initiative, called MicroData Infrastructure (MDI), will allow researchers to directly process firm-level data in various homogenised datasets sourced from different Member States directly. The micro-level dataset, currently covering six countries, is composed of information sourced from a variety of surveys (Community Innovation, the ICT/e-Commerce, and the International Sourcing Survey) and registers (International Trade, Structural Business and Foreign Affiliate Statistics), with potentially more countries and sources to be added in future.

In the same spirit, there is a new open-access, international database that provides a wide range of micro-statistics on income inequality and income dynamics at the individual level. The Global Repository of Income Dynamics (GRID) contains statistics computed from administrative records about earning histories that have been harmonised to be comparable across countries.[4]

GRID currently has data from 13 countries: Argentina, Brazil, Canada, Denmark, France, Germany, Italy, Mexico, Norway, Spain, Sweden, the UK and the US. It planned to expand its coverage to more than 25 countries by early 2025. As explained in detail in Guvenen et al. (2022), GRID's main pillars are that it is based on micro-data sourced from administrative records; it exploits the longitudinal dimension of the data; its granular nature allows for statistics to be produced for finely defined subpopulations; its data harmonisation allows for cross-country comparability. Each country team has produced their respective GRID statistics, and they have written an article about their respective countries with all of the relevant details about the specific databases used.

The OECD has also established a network of researchers working with linked employer-employee data to "enhance our understanding of the role of policies in inclusive growth". The project, called LinkEED 2.0,[5] relies on a network of almost fifty researchers with access to the linked employer-employee data from twenty

[3] See CompNet's dedicated website for further details on this dataset https://www.iwh-halle.de/en/compnet/data/9th-vintage. Last accessed: 1 Feb 2026.

[4] GRID is a joint initiative between the Minnesota Economics Big Data Institute at the University of Minnesota, Princeton University and Stanford University. GRID can be accessed at https://www.grid-database.org. Last accessed: 1 Feb 2026.

[5] https://www.oecd.org/en/about/projects/linkeed-200.html. Last accessed: 1 Feb 2026.

OECD countries. The work for LinkEED 2.0 is structured in five workstreams, namely job displacement in the net-zero transition, gender equality at work, the integration of migrants, job mobility, reallocation and worker careers, and evaluation of structural reforms.

Another relevant case of a cross-country initiative that exploits administrative data can be found in the work of Eurostat, the European Union's statistical office, which is responsible for publishing Europe-wide statistics and indicators that enable comparisons between countries and regions. Eurostat regularly uses administrative data to produce statistics, both as the main data source, and to integrate other types of data.

An example of the first case can be found in the domain of education. To measure progress towards the EU's education and training objectives—e.g. in terms of participation in education and training, learning mobility, education personnel and finance—Eurostat uses a range of different indicators and data sources. Household surveys represent the main source for indicators of educational attainment and adult learning (in particular the EU labour force survey, EU-LFS, complemented by the adult education survey, AES, and the continuing vocational training survey, CVTS). On the other hand, the main data source to monitor other key aspects of formal education systems is the joint UNESCO/OECD/Eurostat (UOE) data collection on education statistics. In particular, UOE provides data on both the participation and completion of education programmes (measured as pupils and students enrolled and new entrants, as well as graduates) and on both the cost and type of resources dedicated to education (i.e. education personnel and education expenditure).

The UOE data collection is entirely based on national administrative sources, with the data compiled by Ministries of Education or National Statistical Offices. To ensure cross-country comparability, this requires a thorough harmonisation process: the participating countries cooperate to develop and apply common definitions, classifications and coverage, which are specified in UOE (2026). This common framework defines what should and should not be included for each type of indicator, such as the number of pupils enrolled in early childhood education programmes. A dedicated annex[6] to the manual defines which programmes fall under the scope of the data collection (i.e. those that can be classified under level 0 of the International Standard Classification of Education).[7]

It should be noted that these definitions might very well differ from the ones applied in the country (to compile national statistics); however, to ensure comparability of international statistics, the UOE data collection relies on the countries' commitments to make all reasonable efforts to report data according to standard international definitions.

[6] https://circabc.europa.eu/sd/a/8402206b-7327-4b40-b894-df3816dfbad2/APPENDIX%20A%20UOE%20manual.pdf. Last accessed: 1 Feb 2026.

[7] https://www.uis.unesco.org/en/methods-and-tools/isced. Last accessed: 1 Feb 2026.

Administrative data can also be used for European-level data collections that are primarily survey based. EU Regulation 2019/1700,[8] establishing a common framework for European statistics relating to persons and households based on data at individual level collected from samples, explains that "to improve data quality and efficiency, the use of administrative records should be encouraged to the extent possible". Article 9 of the Regulation explicitly lists administrative records among the main data sources. The surveys falling under this Regulation include the European Union Labour Force Survey (EU-LFS) and the European Union Statistics on Income and Living Conditions (EU-SILC), for which the rules foresee the possibility of using administrative data to obtain information, provided that the data obtained are of equivalent quality.

As an example, Statistics Finland reports that for the Labour Force Survey, "in addition to interview data, data are obtained from administrative registers available to Statistics Finland from the Register of Completed Education and Degrees and the tax register".[9] A similar situation applies to the European Union Statistics on Income and Living Conditions (EU-SILC). As explained in Lyberg and Lynn (2022), already in the early years of EU-SILC, "in the countries referred to as 'register' countries (Denmark, Finland, Iceland, the Netherlands, Norway, Slovenia and Sweden), most income components and some items of demographic information are obtained through administrative registers". As reported in Wirth and Pforr (2022), as of 2021, "13 countries (Czech Republic, Denmark, Finland, Iceland, Ireland, Latvia, Lithuania, Malta, Netherlands, Norway, Slovenia, Sweden, Switzerland) used register data in combination with survey data for EU-SILC".

Data access is also in the scope of the Common European Data Spaces, created under the European strategy for data of February 2020.[10] Data spaces are designed to facilitate, in a trustworthy and secure manner, data pooling, access, and sharing by institutions, businesses and individuals. Data spaces are meant to offer a clear and practical structure for accessing and using data with fair, transparent, proportionate and non-discriminatory access rules. The European strategy for data advocates for legislative measures to enable access to high-value publicly held datasets across the EU and to allow their reuse for free.

3 Cross-Country Data Linkages

This section discusses the linkage of datasets across national borders. Despite the significant challenges, experience shows that cross-border data linkage is possible, that it enhances data quality and offers evidence for previously unexplored phenomena.

[8] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.LI.2019.261.01.0001.01.ENG. Last accessed: 1 Feb 2026.

[9] https://stat.fi/en/statistics/documentation/tyti. Last accessed: 1 Feb 2026.

[10] https://digital-strategy.ec.europa.eu/en/policies/data-spaces. Last accessed: 1 Feb 2026.

While administrative records generally contain high-quality and comprehensive information, they may lack information about certain cross-border phenomena, such as the level of education attained and jobs held abroad by commuters, migrants and return migrants. This information can be recovered partly using individuals' self-reporting, but it is sometimes completely absent. Cross-border information could be systematically obtained if administrative records from different countries were integrated, so as to allow for regular information exchanges.

Integrating cross-border administrative records is critical for obtaining reliable statistics on individuals' cross-border activities, which is relevant in the EU given that the single market allows citizens to freely move and to work across all 27 Member States. In 2013, 5 million EU citizens aged 15 and over, representing 2.39% of EU labour force, worked in other EU countries in which they had not acquired citizenship. Despite the limitations introduced by the 2020 COVID-19 pandemic, this figure reached 7.3 million in 2025, or 3.33% of the labour force.[11] An increasing number of mobile citizens creates a growing need to integrate administrative records across borders, so as to maintain a representative coverage of the population and accurate information on a diverse set of relevant dimensions, from health records and pension contributions, to tax liabilities and voting rights.

The European Nordic countries are at the forefront of the use of administrative data for research and policy evaluation. The chapters for Finland, Sweden and Denmark in this book offer detailed information about these countries' administrative data coverage and access. Norway also has advanced infrastructure and innovative experience in this field; one initiative, for instance, allows access to microdata even without a formal application.[12]

The Nordic Countries' National Statistical Institutes have been sharing microdata registers with relevant information on cross-border commutes for over two decades; this is probably because the free movement of individuals in the Nordic countries has a long history. For instance, Sweden has been sharing microdata on cross-border commutes with Denmark since 1997 and with Norway since 2001.

However, it was not until very recently that Denmark, Sweden, Finland, Norway and Iceland conducted the first systematic link of their registry data (Brun et al., 2021). In addition to statistics on cross-border mobility in the five Nordic countries, this initiative identified the necessary models and procedures for the exchange and linkage of microdata. Notably, the cross-border linkage of microdata across the Nordic countries allowed for the potential improved accuracy of their national education registers and employment rates that could be realised through regular data exchanges. For instance, the share of individuals aged 15 or over that had their highest level of education updated was 1.3% in Iceland, 0.6% in Denmark and Norway, 0.4% in Sweden, and 0.3% in Finland.[13]

---

[11] Figures for EU-27 Member States from the Labour Force Survey https://ec.europa.eu/eurostat/databrowser/view/LFSA_PGANWS__custom_6625412/default/table?lang=en. Last accessed: 1 Feb 2026.

[12] See https://www.microdata.no/en/. Last accessed: 1 Feb 2026.

[13] See Table 8 in Section 7.3.4 in Brun et al. (2021).

Similarly, the employment rate is underestimated in the 5 Nordic countries, due to the exclusion of cross-border employees in each country's official statistics, ranging from 0.86% in Sweden to 0.03% in Finland. The potential updates to national education and employment records, made possible through data exchange among countries, could significantly enhance the accuracy of research and policy evaluations, particularly given the widespread use of this information as a proxy for individual socioeconomic status.

Another example comes from a group of researchers who has linked register data from Austria and Germany to study the careers of migrants between the two countries.[14] They used a dataset containing information on all workers' working histories in both countries from 1975 until 2014; with the linked datasets, they show how the flows of migrants have been increasing over a considerably long period. They characterise workers who decided to migrate to the other country using their employment history, wages and working conditions (before and after migration) allowing them to investigate the economic returns of migration.

The experiences documented above are clear examples of the benefit brought about by the cross-country linkage of administrative data.

4 Human Interactions

This section reports current cross-border structured human interactions, both before and after conducting data analysis from a causal impact evaluation perspective. It covers some national extra-EU experiences, international organisations and dedicated conference series.

One type of forum of human exchanges is general-purpose scientific conferences in the social sciences. Scientific societies are often de facto international, and they serve as convening fora for researchers to exchange ideas on preliminary results, building new network links and ideas within the science community. The economic, econometric and statistical societies are naturally close to the technical development of the scientific tools for causal analysis; their conferences contain many sessions dedicated to counterfactual impact evaluations.

Some societies are dedicated to the evaluation function, such as the European Evaluation Society, which organises a biennial conference on policy evaluation.[15] Finally, there are conference series dedicated to fostering interactions among the

[14] The researchers are: David Card, Wolfgang Dauth, Wolfgang Frimmel, Julia Schmieder, Andrea Weber and Rudolf Winter-Ebmer. Here is a link to their current project: https://iab.de/en/project/?id=14204276. One of their presentations with name "Selection and careers of migrants between Austria and Germany: Evidence from linked register data" is listed on https://www.uibk.ac.at/en/econstat/research/research-seminars-and-talks/faculty-research-seminar/. Last accessed: 1 Feb 2026

[15] See https://europeanevaluation.org/ees-biennial-conferences/. Last accessed: 1 Feb 2026.

research community, practitioners and policymakers on counterfactual impact evaluations, such as the EU-based "Counterfactual Methods for Policy Impact Evaluation" (COMPIE) conference series.[16]

Other fora for human exchanges are those organised around field communities. Several fields, such as Labour Economics and Development Economics, have recognised the need to integrate administrative records across countries to produce meaningful research. For instance, the Network of European Labour Market Research Institutes (ELMI) was created in 2022 to facilitate the exchange of ideas, data and researchers across its member institutions, with a particular focus on policy advice.[17] ELMI is a network of eleven institutes from nine different European countries studying labour markets and employment. Its members collaborate on research projects from various fields and disciplines of labour market research.

Other cross-country human interactions are facilitated by national and international institutions. The Abdul Latif Jameel Poverty Action Lab (J-PAL) has been a global research network since its foundation in 2003 at MIT in Boston and seeks to reduce poverty by ensuring that policy is informed by scientific evidence, especially via the use of randomised control trials. The outreach of J-PAL's network spans 1,000 researchers at universities around the world.

Recently, J-PAL has recognised the importance of the use of administrative data in the production of high-quality empirical evidence, see Cole et al. (2020). This handbook documents different types of costs associated with the benefit of using administrative data for learning "what works". There is a fixed cost in preparing, cleaning, documenting, linking, and hosting administrative data. There are transaction costs associated with negotiating a Data Use Agreement, which requires considerable time, tact and trust. There are also opportunity costs to pursuing access to administrative data, rather than running a randomised control trial, for instance. Finally, there are also marginal costs associated with repeated queries to a given data set, which can potentially deplete its privacy; the handbook advocates for a so-called privacy loss budget to be fixed in advance.[18]

Other institutions have a mission related to the economic and social improvements in various parts of the world. The Bretton Woods Conference (New Hampshire, US) in July 1944 resulted in the formation of the International Monetary Fund (IMF) and the International Bank for Reconstruction and Development (World Bank). The mandate of the World Bank evolved from post-war international economic cooperation and reconstruction to sustainable development and global partnerships.

The World Bank Group's current President, Ajay Banga, has broadened its mission statement to include knowledge, suggesting to use "not just the financial resources of

[16] https://knowledge4policy.ec.europa.eu/microeconomic-evaluation/compie-conference-series_en.

[17] https://www.elmi-network.eu. Last accessed: 1 Feb 2026.

[18] The issue of how to keep confidential information safe has been conceptualised in the *Five Safes* framework in Desai et al. (2016), which involves the following aspects: (i) safe projects, (ii) safe people, (iii) safe settings, (iv) safe data, (v) safe outputs.

the World Bank to address global poverty, but also the knowledge of member countries to facilitate collaboration on what works and what doesn't", see Semuels (2024). The "knowledge bank" idea appears centred on human interactions, paired with human-data interactions on new data sources. Some of the actions of the World Bank, such as the Development Impact Group,[19] are specifically directed at "generating high-quality and operationally relevant data and research to transform development policy, help reduce extreme poverty, and secure shared prosperity".

Some World Bank case studies discuss human-data interactions; for example, Alix-Garcia et al. (2019) link satellite data to survey data to identify the effects of cash transfers for forest conservation. They also include some policy lessons about data for development in their conclusions, which include making use of data already collected at the household level (administrative data) in place of large surveys with high cost and linking this data to remote sensing-based data (administrative data—extended definition). This is a good example of human-data interaction. Quite naturally, they advocate for the replication of their research design to similar programmes implemented in other countries, which is an example of human interaction.[20]

Another relevant international institution in this context is the Organisation for Economic Co-operation and Development (OECD), an organisation that "works to build better policies for better lives". It was founded approximately 60 years ago, and it aims to "develop evidence-based analysis that helps generate innovative policies and standards to build stronger, more sustainable and more inclusive economies". A recent initiative of the OECD, OECD (2025), focuses on improving the international standard on Public Policy Evaluation.

Another initiative, a joint project between the European Commission and the OECD,[21] fosters the use of counterfactual impact evaluation of Active Labour Market Policies (ALMPs) in European countries and aims to harness the opportunities provided by linked administrative data to support evidence-based policymaking. Apart from demonstrating the power that evaluation brings to policymaking in designing effective and efficient ALMPs, this project has also highlighted the need for high-quality and rich data, linked across registers, see OECD and EC (2025).

The UK "What Works Network" was created in 2013 to ensure that the best available evidence on "what works" is available to the people who make decisions on public services. If evidence is not available, then decision-makers should use high-quality methods to find out what works and this often leads to randomised control trials, see WWC (2018). An integral part of this network is to learn from evidence, so as to guide subsequent experiments and policy design.

In 2017, the US Commission on Evidence-Based Policymaking issued a bipartisan strategy to ensure that rigorous evidence is created efficiently and as a routine part of government operations that can, in turn, be used to construct effective public policy.

---

[19] See https://www.worldbank.org/en/about/unit/unit-dec/impactevaluation. Last accessed: 1 Feb 2026.

[20] See their Section *Data for Development* on page 23.

[21] See https://www.oecd.org/en/about/projects/oecd-ec-project-on-policy-impact-evaluation-through-the-use-of-linked-administrative-and-survey-data.html. Last accessed: 1 Feb 2026.

The US Foundations for Evidence-Based Policymaking Act (Evidence Act) from 2018 is a key step in this direction, requiring federal agencies to develop evidence plans to support policymaking. This has led to the US Federal programme evaluation and the Evaluation Officer Council.

Several other territories in the Commonwealth of Nations have similar approaches to evidence-informed policymaking and rigorous evaluations. Examples include New Zealand's integrated data infrastructure,[22] and the Australian Centre for Evaluation.[23] Within this network and building on this experience, Halpern and Deelan (2024) advocate for better international collaboration on evidence, including funding and sharing of results. This study also includes results of a related survey of senior policymakers in the US, UK, Australia and Canada.

The United Nations (UN) is an international organisation established in 1945 with the primary goal of promoting peace, security and cooperation among its 193 Member States. The UN's policy evaluation work is a critical component of its efforts to achieve its goals, and it is carried out through various agencies, such as the United Nations Development Programme (UNDP), the World Health Organisation (WHO), the United Nations Children's Fund (UNICEF) and the Food and Agriculture Organisation (FAO). These agencies conduct evaluations to assess the effectiveness and impact of UN programmes and policies, and also employ causal methodologies.[24]

The UN's policy evaluation work has also led to the development of new guidelines and standards for evaluation, such as the UN Evaluation Group's (UNEG) "Norms and Standards for Evaluation", which provide a framework for evaluating UN programmes and policies, and promote a culture of evaluation and learning across the organisation by fostering human interactions. For example, the UNDP's Independent Evaluation Office conducts evaluations of its development programmes, focusing on areas such as poverty reduction, governance and environmental sustainability.

Similarly, the WHO's Evaluation Office assesses the effectiveness of its health programmes, including those related to infectious disease control, nutrition and strengthening health systems. FAO's Office of Evaluation also plays a crucial role in policy evaluation, particularly in the areas of agriculture, food security and rural development. For instance, evaluations of FAO's programmes to support small-scale farmers have informed the development of more effective strategies to improve agricultural productivity and reduce poverty.

---

[22] See https://www.stats.govt.nz/integrated-data/using-microdata-for-research/. Last accessed: 1 Feb 2026.

[23] https://evaluation.treasury.gov.au/about/about-australian-centre-evaluation. Last accessed: 1 Feb 2026. See also https://www.canada.ca/en/treasury-board-secretariat/services/audit-evaluation/evaluation-government-canada.html. Last accessed: 1 Feb 2026.

[24] The FAO Office of Evaluation (OED) "generates evidence and recommendations that inform thinking, deliberations and, ultimately, decisions of FAO" and its stakeholders. Recently, it organised the "Advancing Evidence-Based Decision-Making: GCF Project Evaluations, workshop 2025", see https://www.fao.org/evaluation/resources/events/detail/advancing-evidence-based-decision-making--gcf-project-evaluations-workshop-2025/en, "to exchange insights, align evaluation strategies, and enhance the impact of Green Climate Fund evaluations".

5 Support Measures

This section discusses possible support measures aimed at improving cross-border exchanges on learning what works.

Policy learning requires popular consensus about its ability to improve the outcomes of public policies for the general public. Activities to discuss the use of administrative data and the related evidence on policies are part of the opportunity set for policymakers. Policy learning should be presented as a way to improve the public sector's accountability and transparency and to improve its effectiveness. These elements can improve the level of trust in the public sector, which appears all the more necessary in the current public discourse, see OECD (2025).

Another possible policy action is to foster harmonised research design in policy evaluations across countries. Bertheau et al. (2023) provide an example of a common design project on the effects of job loss in different countries. They investigate the effects of losing a job in Austria, Denmark, France, Italy, Portugal, Spain and Sweden. They adopt a common research design, a common definition of displacement event and a common criteria for sample selection; this allows them to compare estimates on the consequences of job displacement across countries.

Importantly, this research allows for a comparison of the effectiveness of mitigating measures in the different sets of active labour market policies implemented in each country. With such a harmonised research design, one can answer the "what works" question with a first step towards understanding "why" the same policy may work differently in different countries.

A different and related support to cross-country policy learning comes from the use of the results of causal impact evaluation to improve the policy effectiveness and its implementation. An example comes from the use of recurrent evaluation on the effectiveness of training programmes in Estonia to guide the choice of which course to recommend to the unemployed, see Chap. 9 in this book on Estonia. A different related benefit concerns the use of evidence from one country in the design of a similar policy in another country.[25]

There are several enabling circumstances and possible policy interventions that might help to foster these interactions: communication about the public value of policy learning can be supported by specific funding; funding and infrastructure can help data linkage and data interactions; completed projects can help demonstrate the value added generated by policy learning; harmonised research designs can help learning differential effects of policies, and this could be supported by specific funding.

[25] Including a proper discussion of issues of external validity.

6 Conclusions

Human and data interaction can greatly improve researchers' productivity, as well as improving trust in the public sector and its effectiveness. Moreover, learning about what works can directly link to better services for the public. This is a win-win situation.

One key driver of improvement in the creation of public value associated with policy learning is improved interactions concerning data, including harmonisation, matching, triangulation and the interoperability linkage of different administrative data sources. As discussed in this chapter, some networks and institutions favour and support such interactions; greater coordination among them could imply higher synergies and returns on investment.

Another key driver is improved human interactions on research designs, identification strategies, learning from previous analysis, and the existence and accessibility of data. These types of activities are now happening under different institutional settings; also here, greater coordination could also mean higher synergies.

An improved public discourse on the public value of policy learning might help coordination and increase the public taste for supporting the use of administrative data for policy learning; this chapter and this book attempt to contribute to this last strand.

References

Alix-Garcia, J., Sims, K. R. E., Orozco-Olvera, V., Costica, L., Fernandez Medina, J. D., Romo-Monroy, S., & Pagiola, S. (2019). *Can environmental cash transfers reduce deforestation and improve social outcomes? A regression discontinuity analysis of Mexico's National Program (2011–2014).* World Bank Policy Research Working Paper.

Bertheau, A., Acabbi, E. M., Barceló, C., Gulyas, A., Lombardi, S., & Saggio, R. (2023). The unequal consequences of job loss across countries. *American Economic Review: Insights, 5*(3), 393–408.

Brun, N., Ekmark, S., Munch Haagensen, K., Hardarson, O., Rustad Holseter, A. M., Nome Næsheim, H., & Ruotsalainen, K. (2021). *Nordic cross-border statistics: The results of the Nordic Mobility project 2016–2020.*

Carattini, S., Dur, R., & List, J. (2024). Policy evaluation and the causal analysis of public support. *Science, 386*(6721), 490–492.

Cole, S., Dhaliwal, I., Sautmann, A., & Vilhuber, L. (2020). *Handbook on using administrative data for research and evidence-based policy.* J-PAL.

Desai, T., Ritchie, F., & Welpton, R. (2016). *Five safes: Designing data access for research.*

EC. (2021). *Better regulation guidelines* SWD(2021) 305 final. European Commission.

Guvenen, F., Pistaferri, L., & Violante, G. L. (2022). Global trends in income inequality and income dynamics: New insights from GRID. *Quantitative Economics, 13*(4), 1321–1360.
Halpern, D., & Deelan, M. (2024). *Global evidence report*. Economic and Social Research Council.
Lyberg, L., & Lynn, P. (2022). *Improving the measurement of poverty and social exclusion in Europe: Reducing non-sampling errors*. Publications Office of the European Union.
Nagaraj, A., & Tranchero, M. (2023). *How does data access shape science? The impact of federal statistical research data centers on economics research.* National Bureau of Economic Research. 31372 Working Paper.
OECD. (2025). *OECD recommendation on public policy evaluation.*
OECD and EC. (2025). *Counterfactual impact evaluations of active labour market policies: Lessons from using linked administrative data.* OECD Publishing, Paris.
Semuels, A. (2024). Ajay Banga wants the World Bank to work as a "knowledge bank" to reduce poverty. *TIME Magazine*.
UOE. (2026). *UNESCO-UIS / OECD / EUROSTAT manual on concepts, definitions, and classifications.* Guidelines Joint UNESCO-UIS / OECD / EUROSTAT Data Collection Education. Version 2026. Montréal, Paris, Luxembourg 2026.
Wirth, H., & Pforr, K. (2022). The European Union statistics on income and living conditions after 15 years. *European Sociological Review, 38*(5), 832–848.
WWC. (2018). *The what works network: Five years on.* UK Government.

Josep Amer-Mestre is a Quantitative Policy Analyst at the Competence Centre on Microeconomic Evaluation (CC-ME) of the Joint Research Centre of the European Commission. His research focuses on evaluating public policies in the fields of education and the labour market. Josep holds a Ph.D. in Economics from the European University Institute (EUI).

Sara Flisi is a Scientific Officer at the Competence Centre on Microeconomic Evaluation (CC-ME) of the Joint Research Centre of the European Commission. Her main research interests lie in the areas of impact evaluation, economics of education and labour economics. She holds a Ph.D. in Labour Economics from the University of Modena, and a MSc in Economics and Econometrics from the University of Nottingham.

Paolo Paruolo is an econometrician, currently Head of the European Commission Competence Centre on Microeconomic Evaluation since 2016. He holds a Ph.D. in Econometrics from the University of Copenhagen, and he has been Full Professor of Econometrics at the University of Insubria, Varese, Italy, and previously at the University of Bologna. His research focuses on econometric theory, time series analysis and counterfactual methods. He has published in *Econometrica*, *Journal of Econometrics, Econometrics Journal*, *Econometric Theory*, *Journal of the Royal Statistical Society Series A* and *Oxford Bulletin of Economics and Statistics*, among others. He has written and edited several books, including "Data-driven policy impact evaluation: How access to microdata is transforming policy design", Crato, Paruolo (eds 2019).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 30
Harnessing Administrative Data for Policy Evaluation in Labour, Education and Health Economics



Nuno Crato, Cinzia Di Novi, and Elena Claudia Meroni

Abstract This chapter highlights the importance of individual registry data when evaluating policy effectiveness in labour, education and health economics. Administrative data often offers a more accurate and reliable alternative to survey data, enabling researchers to track individuals over time and to assess the impact of various policies. Policymakers can identify areas of success, improvement and further growth by leveraging registry data, ultimately maximising the impact of their investments. The use of administrative data has been instrumental in addressing complex research questions, particularly during the COVID-19 pandemic. Effective policy evaluation relies on the integration of various registries and on the creation of a comprehensive database that facilitates cross-field research. This approach enables researchers to explore causal pathways and to provide more precise assessments of policy effectiveness. Ultimately, the strategic use of administrative data can inform decision-making and can drive evidence-based policy reforms in labour, education and health economics. Researchers and policymakers can work together to create more effective and targeted interventions by harnessing the power of registry data.

1 Introduction

This chapter will cover three principal areas in which individual data can be both collected and used for research purposes (i.e., labour, education and health economics). Individual registry data is often collected through administrative

N. Crato
University of Lisbon, Lisbon, Portugal
e-mail: ncrato@iseg.ulisboa.pt

C. Di Novi
University of Pavia, Pavia, Italy
e-mail: cinzia.dinovi@unipv.it

E. C. Meroni (✉)
European Commission Joint Research Centre, Ispra, Italy
e-mail: elena.meroni@ec.europa.eu

© The European Union and The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_30

processes, such as by using medical records, tax records, unemployment spells, student enrolment or attendance records; these data can also be combined with other data sources to provide a more comprehensive understanding of various sectors and policies. A brief description of the data and its main uses will be provided for each of these three research areas, with a specific focus on the advantages of using them to evaluate the effectiveness of policies. Indeed, assessing the effectiveness of EU (European Union) or national policies (e.g., the introduction of a minimum wage, a change in the educational system or a healthcare system reform) or of specific interventions (e.g., Active Labour Market Policies—ALMP, interventions in schools aimed at boosting students' achievement or screening programmes, vaccination and health promotion interventions) is crucial to determine whether the resources invested have achieved the desired objectives, such as by increasing employment rates or by improving students' school performance or individuals' health status.

Numerous EU initiatives both fund and encourage activities within these domains. For instance, the European Pillar of Social Rights emphasises the necessity to foster employment creation and equipping individuals with the necessary skills for these jobs, along with ensuring equal access to improved healthcare. Similarly, NextGenerationEU and its main component, the Recovery and Resilience Facility, established in response to the COVID-19 pandemic, support investments in employment, social protection, education and healthcare.

Determining whether the implemented actions yielded the desired outcomes is extremely important: it helps to identify areas of success, areas for improvement and opportunities for further growth, ultimately ensuring the greatest possible impact from EU investments. Counterfactual Impact Evaluation (CIE) methods provide the means by which to evaluate the effectiveness of such policies or interventions: these methods allow for estimating the true causal effect by predicting the outcomes for participants in the absence of the programme. The use of CIE methods requires possession of information about a potential control group, such as a group of individuals who did not receive the treatment (or who were not subject to the policy), and having information for both the treated and the control group over time. These characteristics make administrative data particularly suitable for conducting CIE in the fields of labour, education or health economics (See, e.g., Crato, 2023, p. 128 for a comparison of the types of data useful for CIE studies).

The remainder of this chapter is structured as follows: An overview of the administrative data, its features, and its advantages compared to survey data will be presented for each of the three areas—employment and labour, education and health. Subsequently, illustrative examples of how this data might be employed in research will be provided. Lastly, the chapter will highlight how this data was utilised to advance knowledge during the COVID-19 pandemic, thereby demonstrating its value in comprehending both the pandemic's short-term and long-term effects.

2 Employment and Labour

2.1 Overview of Data

To gain deeper insights into the field of labour economics, it is first essential to have data on employment status, job positions and wage levels of individuals. Additionally, details about employers are valuable, including the sector type, company size and geographical location. There may also be an interest in the duration of an individual's unemployment, as well as regarding the nature of any assistance that they received while out of work.

One of the most reliable sources of information is either registry or administrative data. Labour economics is among the pioneering fields in which researchers have begun utilising administrative or registry data (Heckman et al., 1999). The first countries that relied on such data for statistical purposes were the northern European countries (Denmark, Finland, Iceland, Norway and Sweden) as well as the UK and USA. The accessibility of registry data enables the investigation of policy effectiveness and informs policymakers about what works and why.

Administrative data are certainly more efficient for evaluation purposes than ad hoc surveys because they offer more accurate and reliable information on individuals' employment and financial histories, thereby enabling individuals to be tracked over time, both before and after an intervention. Survey data are often limited by low response rates, recall bias and inaccurate reporting of sensitive information, such as income. Using registry data for evaluation purposes allows researchers to conduct more robust analyses, thereby leading to better-informed policy decisions.

In particular, registry data offer more accurate retrospective information when compared to surveys, given that it is challenging for individuals to recall past employment or unemployment spells. As shown by Sánchez et al. (2014), in a study comparing survey and registry data in Sweden, individuals tend to wrongly report the start and end day of past jobs and total number of unemployment spells, especially when the recall time is large (only 54% of subjects reported correct number of spells of unemployment when the recall time was 15 months, with that figure plummeting to 7.5% when the recall time was extended to 11 years).

Additionally, when the objective is to evaluate the impact of a policy intervention, registry data provides precise and reliable information on both treated and non-treated individuals, which is essential for accurate impact studies, especially when using wages or income as the outcome of the evaluated intervention. In particular, income-related inquiries in surveys often go unanswered (as seen in studies like Yan & Curtin, 2010). Such sensitive questions might be viewed as an invasion of privacy by respondents (Tourangeau & Yan, 2007, p. 86). Additionally, it is argued that these questions can either be intimidating or demanding for survey participants (Loosveldt & Storms, 2008) due to challenges in comprehending the concepts or terminology in financial questions or difficulties in retrieving the required information to provide accurate responses (Moore et al., 2000).

Finally, administrative data can allow for the linking of individual employment data to firm data, in a Linked Employer-Employee Dataset (LEED) settings. Usually, such datasets are comprised of a person file, a job file and an employer file. Employed people are linked to employers via jobs. A person can have a number of jobs throughout the year with one or many employers, some of which may be held concurrently. LEED data can provide a wealth of information about job mobility, career progression and can allow for more accurate wage analyses, given that it is possible to control for factors such as firm size or industry.

Administrative data sources typically include social security, fiscal files, insurance records, tax records, archives from the public employment services and pension records. The following individual level information is usually found in administrative data on employment: demographic information (date of birth, gender, ethnicity), and education level; employment status (employed, unemployed, or not in the labour force) which is usually recorded in spells, with starting and ending dates for each employment/unemployment spell; hours worked (weekly hours worked); wages and salaries (hourly wages, weekly earnings and annual salaries); job training and education or participation to other active labour market policies (information on the types of training and education programmes offered during unemployment and/or employment); job title (the specific title or position held by the employee) and geographic location (country, region or city of employment).

It is also possible to know the employment tenure when linked employer-employee data is available, namely the length of time that an individual has been with their current employer and the employer characteristics, specifically industry, size and employer's location.

2.2 Examples of Use

Administrative data have been used extensively in labour economics research. In this section, their use will be discussed in two explicative examples: the evaluation of ALMPs and the impact of job displacement.

ALMPs hold significant importance within the European Employment Strategy (EES), which regards employment as a primary goal. The primary categories of ALMPs in European countries consist of training programmes, private sector incentive schemes, direct employment programmes and Services and Sanctions and cover measures aimed at improving job search efficiency, such as counselling, monitoring or job search assistance. Targeted ALMPs are addressed to young job seekers (aged 25 and below) who are unemployed, as well as individuals with disabilities, in several European nations. The primary financial instrument employed by the European Commission (EC) to support these initiatives is the European Social Fund (ESF), which seeks to both foster and promote employment, among other objectives, and it is one of main instrument by which to support the implementation of the European Pillar of Social Rights.

Administrative data is now standard when evaluating ALMPs. A recent study by the European Commission, a meta-analysis focusing exclusively on ESF-funded active labour market policies, found that 766 estimates came from studies using administrative data, while only 77 came from studies using survey data (European Commission, 2023, Table 4). Abundant evidence exists on the effectiveness of ALMPs in Europe and worldwide thanks to the utilisation of administrative data in evaluation. The meta-analysis reveals that European Social Fund (ESF)/Youth Employment Initiative (YEI) interventions within the European Union are effective, with an average estimated effect size of a 7.3 percentage point increase in the employment probability of treated individuals.

Card et al. (2010) also showed that 85% of the studies rely on administrative data (Table 4) in another meta-analysis covering all ALMPs globally and this involved the use of two main approaches. The first one relied on longitudinal administrative data on employment and/or earnings for the participants and for a comparison group (who are assigned to a simulated starting date for a potential programme). Typically, the data set includes several years of pre-programme labour market history and propensity-score matching was also used to narrow the comparison group to a sample whose observed characteristics and pre-programme outcomes closely matched those of the participants. The second one used a duration model of the time to exit from registered unemployment since all the necessary data can be drawn from the benefit system itself in many countries (particularly the Nordic ones) without the need access to employment records. In the follow-up study, Card et al. (2018) also showed that most of the evidence comes from the same methodological approaches, which implies the use of registry data.

The impact of job displacement on labour market outcomes, such as employment, wages and income, has been the subject of extensive research; several studies exist that use single countries' registry data. However, the lack of a consistent methodology and harmonised data across countries has made it challenging to compare the labour market consequences of job displacement across different nations. However, a recent study by Bertheau et al. (2023) has overcome this issue, thereby laying the groundwork for future harmonised research. Their paper built and analysed a harmonised dataset that combines matched employer-employee administrative registers from seven countries over the course of three decades in Austria, Denmark, France, Italy, Portugal, Spain and Sweden. They focus on assessing the labour market effects of job displacements, defined as the permanent loss of a long-term job due to mass layoffs or establishment shutdowns for economic reasons. This work provides comparable estimates about the labour market consequences of job displacement across countries by adopting a common research design, a common definition of the displacement event and by using identical sample selection criteria. The results show that the picture that emerges is quite different from the comparison of findings obtained in studies focusing on the single countries. This highlights the importance of using a harmonised research design and harmonised data when conducting a cross-country analysis. The study was conducted in countries with accessible administrative data, allowing information to be traced back for a reasonable number of years. Further research of this

kind is encouraged and should emphasise the importance of administrative data in such endeavours.

2.3 Administrative Data in Labour Economics Research and COVID-19 Pandemic

The COVID-19 pandemic brought an enormous shock to labour markets worldwide. Entire sectors, such as tourism and hospitality, were forced to cease operations, while many others experienced a drastic reduction in working hours. Workers were forced to work from home where possible. Extensive research was conducted on the effect of COVID-19 on several labour market outcomes during the pandemic's initial months. It should be acknowledged that many ad hoc surveys were conducted during those critical months, and several papers used existing surveys (Current Population Survey—CPS, Labour Force Survey—LFS). However, and where possible, the use of administrative data was helpful in providing evidence about the pandemic's impact.[1]

For example, Bussink et al. (2022) used registry data from Centraal Bureau voor de Statistiek (CBS) from The Netherlands to study the impact of the pandemic on graduates' prospects in terms of employment probabilities. Relying on the same data source, Hassink et al. (2020) used weekly administrative panel microdata from Dutch employees to study whether individual labour market outcomes, as measured by employment, working hours and hourly wages, were more strongly affected in provinces where COVID-19 confirmed cases, hospitalisations and mortality were relatively high. Casarico and Lattanzio (2022) estimated the impact of firing freezes and easing the requirements to access short-time work compensation schemes in Italy by relying on administrative data on mandatory notifications that firms submit to relevant public agencies and to the Ministry of Labour and Social Policy.[2] Cuntz and Sahli (2022) used data on public insurance scheme records for artists located in Germany to estimate the wage loss of various categories of artists during 2020. Gulyas and Pytka (2020) use the universe of Austrian unemployment insurance records, until May 2020, to document that the composition of unemployment insurance claimants during the COVID-19 outbreak was substantially different compared to previous times. Nunes et al. (2023) study the impact of the pandemic crisis on jobless claims and job placements by using monthly data covering the universe of individuals registered as unemployed Portugal (from the Portuguese Public Employment Service), complemented with electronic payments, linked employer–employee

[1] The references reported refer to the final publications in peer-review journals; however, all of the papers were available as working papers already in 2020.

[2] They have information on the exact start date and, if the contract is terminated, on the end date and the reason for its termination (mainly, firings, termination of temporary contracts and voluntary quits) on the type of contract (open-ended or temporary, fulltime or part-time), detailed occupational and sectoral codes (6-digit Isco and Nace, respectively) and individual characteristics of workers, such as gender, year of birth, the region of domicile and work and their education level for each contract.

data, and furlough records. Campa et al. (2021) study gender differences in the effect of the pandemic by using aggregated monthly data with population-wide coverage from the registry of unemployed that is collected and kept by the Swedish Public Employment Service, complemented with information on the average wage by occupational group and gender as reported by Medlingsinstitutet and publicly available at Statistics Sweden. Hensvik et al. (2021) provide empirical evidence about the impact of the COVID-19 crisis on job searches by using job ads posted and the search activity on Sweden's largest job board (Platsbanken), which is operated by the Swedish Public Employment Service (PES).[3] Adermon et al. (2022) use monthly administrative data on labour earnings, social insurance benefits and means-tested income transfers for the universe of the Swedish working age population to assess how the COVID-19 pandemic affected labour earnings across the income distribution and for various groups on the labour market.

It should be noted that the countries that are used to employ extensively administrative data for research and statistical reporting, such as Sweden or The Netherlands, could also provide quicker evidence about the effect of COVID-19 by relying on such registry data.

3 Education

3.1 Overview of Data

When assessing educational policies, key metrics typically include what we may call *quantitative* data, such as enrolment, graduation and dropout rates, and what we may call *quality* data (i.e., student cognitive and skills performances as measured by assessment scores across different academic grades). This distinction is somehow arbitrary, but it has the merit of highlighting the fact that it is one thing to enrol, to remain in school and to make students proceed across school grades, but it is quite another to succeed in increasing students' quality of education. Traditionally, studies on educational systems concentrate on what we have just called quantitative data. Recent research and international studies such as PISA (Programme for International Student Assessment), TIMSS (Trends in International Mathematics and Science Study) and PIRLS (Progress in International Reading Literacy Study) have helped to increase researchers and policy analysts' attention to the cognitive variables.

For quantitative data, registries compiled by government bodies or educational institutions, being administrative in nature, provide a wealth of information that is useful to assess the effectiveness of policies designed to reduce school dropout rates, to boost higher education enrolment or similar undertakings. Furthermore, if well collected, these data enable the longitudinal tracking of students, thereby offering

[3] Firms can post vacancies and screen applicants on Platsbanken.

a comprehensive view of their academic career from primary through to higher education levels.

There is a plethora of complex instruments for the analysis of the quality of education (i.e., for the analysis of student's performance). It is a complicated structure with many interconnected and mutually reinforcing components.

There is a bottom-up movement, starting with formative assessments by teachers during normal classes, which prepares students for summative assessments done by teachers and schools at the end of the term or of the school year. These two types of *internal* assessments prepare students for national exams, which are *external* summative assessments, are typically standardised (i.e., uniform) and that have uniform grading. Both of these assessments can be considered as administrative data and are usually recorded in registries, thereby also allowing a comparison of teachers' vs standardised assessments. (Examples here include either the Italian INVALSI or the Dutch data from CBS, both of which contain both national standardised and teachers' grades) There is also a top-down effect because these national exams set the level of rigour and the tone for tests in schools and for daily activity in classes. They calibrate the levels of ambition in schools' teaching activities and schools' internal assessments (Koretz, 2008, p. 23).

While internal assessment activities cover all or almost all school subjects or areas of student life, external standardised tests scores evaluate pupils' proficiency in particular subjects and particular academic levels. Typically, these assessments focus on knowledge and skills in mathematics and in the national language and may even be extended to science and foreign languages. They are held at the end of school cycles such as elementary, middle and high school levels.

There is also another level of assessment, the international one. We are talking about the so-called International Large-Scale Assessments, or ILSA, of which the main examples are the aforementioned PISA, TIMSS and PIRLS surveys. These international assessments started in 1995, 2000 and 2001, respectively, but are already invaluable for benchmarking a nation's performance against others and for tracking its progress over time. However, these are just surveys; they are not exhaustive and do not provide data at a granular level. They are conducted every three, four and five years, respectively, and they sample only a small fraction of the student population; they also focus on specific age groups or grades—for instance, PISA assesses 15-year-olds exclusively, and TIMSS and PIRLS assess 4th-graders only.

The release of these international results was a wake-up call for their relative situation for many countries and provided valuable help to giving more prominence to statistical studies and policy measures centred on the quality of education offered.

3.2 Examples of Use

There are numerous examples of statistical and counterfactual studies in the field of the economics of education. Their impact on education policy varies widely, ranging from highly influential studies that inform and shape policy decisions, to those that

have a moderate impact, influencing specific aspects of education reform, and finally to studies that, despite their potential, remain largely overlooked or underutilised, failing to translate into meaningful policy changes. A very successful example is the generalisation of pre-school education. Most countries either started or reinforced early-school programmes and this practice is no longer a rarity following seminal studies by Heckman, Cunha, Carneiro and others (see, e.g., Carneiro & Heckman, 2002) which showed the positive impact in equity of pre-school education.

A very important example of a consensus, both regarding the literature and the public political stances, is the relationship between students' results and their socioeconomic status. Administrative data has repeatedly shown a persistent correlation between these two variables, and this has prompted the discussion of equity policies. The problem persists, unfortunately, and many policies adopted in the name of inequality reduction have had perverse effects. What we can say is that the issue garnered some attention at least. We direct the reader to a modern and authoritative survey because the literature on these topics is immense (Blanden et al., 2023).

Another example is the teacher's value-added issue. Many studies with administrative data have shown a large variability of students' results attributed to the teachers (e.g., Chetty et al., 2014; Hanushek, 2011). As Koedel et al. (2015) have put it, there is a "consistent finding in research that teachers vary dramatically in their effectiveness." However, as they also write, its "application in public policy, and in particular the use of value-added to help inform teacher evaluations" has been very controversial and limited. This case is very interesting because the impact of teachers' quality on the students' results is obviously very significative, but the fierce opposition from unions and other political forces have practically prevented actions in an area that could improve youngsters' education.

Examples in this vein abound, from the dubious but widely overblown impact of public investment (Crato, 2021, pp. 7–9), to the marginally useful but widely inflated class size reduction (Ecalle et al., 2019). This leads some researchers to conclude that education is often a frustrating area regarding the impact of evidence on policy measures (Gomendio, 2021).

The recent "reading wars" are an extreme negative example. These "wars" confront two opposing views on early reading. On the one side, we find supporters of a systematic phonics approach, developing fluency in decoding text, starting from syllables to words and to sentences, and from these bases leading children to advance to reading comprehension. On the other side, we find supporters of some form of learning that avoids systematic decoding and promotes whole words or whole phrases reading, based on cues from the global sense of the text or even illustrations. "Whole language" and "balanced literacy" are examples of these latter methods (Castles et al., 2018). Although statistical data analysis and educational psychology scientific evidence has been overwhelmingly supporting phonics instruction for initial reading learning, many countries and regions persisted, and some continue to persist in other faulty reading instruction methods (Buckingham, 2019; Castles et al., 2018; Chall, 1967; Cook et al., 2017; Lane et al., 2025).

3.3 Administrative Data in Education Research and COVID-19 Pandemic Effects

The recent COVID-19 pandemic was arguably the most serious disruption event in recent times. It had implications in all aspects of social life. One of these aspects included education. Years later, we estimate that students worldwide had learning losses of between 15 and 30% of a standard deviation of national standardised results (Jakubowski et al., 2023, 2024). These learning losses have deep implications for the life of the generations involved, for students' future earnings and for the economic future of our societies (Hanushek & Woessmann, 2020). The consequences of the pandemic and the measures adopted to counteract them are now very important topics of research. A few researchers were able to use administrative data to assess the extent of the problems, even during the COVID-19 disruption. One of the earliest examples is Maldonado and De Witte (2022). This work, remarkably, appeared in working paper format six months after the first school closures, which happened in mid-March 2020. These researchers were able to estimate the learning losses in the first year of school closures by studying the results of Flemish standardised tests and a large administrative data set. Data also allowed them to detect differences in losses between disciplinary subjects and socioeconomic backgrounds.

At about the same time, Engzell et al. (2022) were able to detect similar effects in the neighbouring Netherlands. They also used standardised examinations and administrative data and were able to detect similar losses and asymmetries.

After the pandemic was over, we came to know much more. We now know that country or region average number of days of school closures was correlated with learning losses as measured by PIRLS 2021 and PISA 2022, the two international post-pandemic assessments held up to today (Summer 2024). We also now know that there are cases of no or almost no school closures and significant learning losses. There were also countries with a similar number of days of school closures and wide differences in losses. Interestingly, some countries and regions, namely Singapore and Taiwan, were even able to improve their results after suffering pandemic restrictions. These cases are remarkable, but we are not aware of any serious interest from other countries in learning about these experiences.

Even as the learning losses are now known and widely recognised in the academic world, many policymakers only pay lip service to these findings and act as if this were not a serious problem in need of a precise evaluation and for consequent action. Countries' approaches varied in many aspects; namely in the seriousness of education measures, in the strength of the student assessment and in the disclosure of data. We will now briefly discuss each of these three aspects.

First, regarding education measures, most countries adopted one form or another of restricting school social contacts. Only a few of them did not close schools for any period. These data are available in regional or country aggregates, but data are very coarse, and it is difficult to draw conclusions for those cases in which there have been school differences in education policies within a country or region.

Second, regarding assessment, it is not surprising that the international experience during and after the pandemic has highlighted the importance of valid student assessment. Some countries, such as The Netherlands, maintained regular mid-year standardised low-stakes national testing during the pandemic period, and this has enabled them to understand the problems that they face.

Other countries decided to suspend national exams and other standardised tests, but then resumed the assessment procedures soon thereafter; these countries were not surprised by the negative PISA results. A good example is the US: they maintained their National Assessment of Educational Progress (NAEP) tests and later witnessed downward movements of PISA parallel to NAEP scores.

Another group of countries, namely Spain and Portugal (Gomendio, 2025; Marôco, 2024), had already changed to poor systems of assessment a few years before, and additionally suspended some of the standardised tests during the pandemic. It is not yet clear whether these countries' governments are going to resume regular testing or are going to see if this break worked as an opportunity to avoid assessment.

This issue is profoundly serious because these and other countries are experiencing a double blow: the decline in learning due to the pandemic losses adds to a decline in learning due to curriculum, other education policy issues or social problems more generally. As various researchers have highlighted (see various examples in Crato & Patrinos, 2024), there is an increased risk that faulty education policies will persist, leading to further deterioration of the education system, in the absence of clear assessments that can disentangle the complicated causes of learning losses. Regular, frequent and standardised student assessments are needed now more than ever—and results need to be known. We cannot act on what we ignore.

Even in countries in which national exams have continued, the results can be biased, as reported by Herrera-Pavo et al. (2024) for Ecuador. This bias can reach serious proportions, as reported by Marôco (2024) for Portugal. Internal assessments organised by Portugal's Ministry of Education were systematically biased upwards, in such a way that they seemed to show that students' knowledge and skills had improved with school closures. For Portugal and other countries, the disclosure of PIRLS and PISA results was a shock.

In Europe, the only country that has consistently shown excellent results on PISA surveys is Estonia. It is an example that deserves to be studied. This country's education has also suffered from the pandemic, but it maintained a policy of high expectations and rigorous external assessments. As this country expert in assessment explains (Tire, 2024), Estonia considers their national exams to be instruments through which to improve student results. This way, changes were seen as temporary when changing national tests from high to low stakes during the pandemic. High-stakes tests were reinstated after the pandemic. Interestingly, it was observed that students reduced their efforts during the low-stakes exams period.

Third, the disclosure of students' assessment is vital for researchers and for society at large. To obtain, study and make students' results available is not enough. Most recently, a new (and old) obstacle remains paramount: we need timely reported results.

Most schools, governments, teachers and families make their most important results fall at the end of the school year. If results arrive too late, then they are useless for many purposes. There is no longer time for policymakers to correct faulty policies or for schools, teachers and families to adjust to the situation.

Timely policy impact evaluation should also be part of countries' collective efforts. Substantial financial efforts should be accompanied by evaluations of the results. In the US, for instance, massive amounts of funding have gone to schools through various pandemic-relief funds totalling c. $200 billion dollars. However, the results of these investments, and the extent of the hoped-for recovery of test scores in general, remains unclear (Hanushek & Strauss, 2024). As a result, it may be difficult to point to any approach to recovery that has worked better.

All of this provides a forceful argument for timely student assessment and for timely results disclosure. The bases for policy evaluation are hindered if this assessment is not obtained and known. In a way, student assessment is the most basic data to be collected. Policy evaluation is blind if we do not follow this variable. The progress of students is the education policy's most important outcome.

4 Health

4.1 Overview of Data

The primary objective of policy analysis in the healthcare sector is to investigate the impact of an intervention on patients, healthcare professionals or hospitals. Concerning labour or educational data, health information can be collected through surveys or registry data. The following sections will particularly focus on administrative health data concerning patients for whom there is also a greater availability of information and the possibility of making comparisons with data collected through surveys, whose use is spread widely among researchers involved in health economics.

Administrative data are routinely collected whenever a patient interacts with healthcare services. These data encompass a wide range of information including screenings, surveillance, existing health conditions, diagnoses, treatments, patient outcomes, laboratory test results, billing data, reimbursement claims as well as patient-generated data from wearable devices and mobile apps. Stored in individual datasets, administrative healthcare data can be linked together using a data set identifier assigned to each patient, replacing their real-world identifier. This data set identifier facilitates the identification of multiple records within the data set that relate to the same individual. For example, it aids in detecting re-hospitalisations or mortality events and enables linkage with other administrative records to provide a more comprehensive understanding than any single data set could offer alone.

Administrative hospital discharge data are commonly utilised in health economics and in healthcare policy evaluation research. This type of data has been employed

extensively by health services and population health researchers to investigate healthcare outcomes, effectiveness, appropriateness and utilisation of healthcare services. Additionally, it is used to study population health status and its determinants. Several characteristics of this kind of administrative data have contributed to their extensive use. For instance, they contain detailed information about hospitalised patients, such as age, health status (including the presence of multiple chronic conditions and specific chronic diseases, as administrative hospital discharge data typically include a detailed medical history) and socio-demographic information (e.g., gender, nationality, place of residence and marital status). Additionally, patients can be readily identified within the sample by using the International Classification of Diseases (ICD), the standard system for assigning codes to diagnoses and procedures related to hospital utilisation that has been adopted in many countries. Researchers can establish extraction criteria based on these codes and their placement in discharge abstracts, thereby focusing on primary diagnoses. This method enables researchers to select patients of interest based on specific medical conditions (Quan et al., 2004). For instance, a criterion for identifying cancer patients could involve detecting a C00-D48 code listed as the primary diagnosis.

However, it is important to consider that administrative data, such as hospital discharge records, may have limitations when used for research purposes. Firstly, errors can arise during data creation due to inaccurate diagnoses by physicians, incomplete documentation of clinical information in hospital records or errors in coding diagnoses and procedures by administrative staff. Additionally, administrative data, which are not collected specifically for research purposes to evaluate policy interventions or answer other research questions, often lack detailed patient information. This information could include important confounding factors that influence health outcomes, such as the patient's overall health status, access to healthcare services and mortality. Administrative data, such as hospital discharge records, for instance, often do not contain detailed information on lifestyle unless the patient has been hospitalised for specific conditions related to lifestyle itself (e.g., smoking habits in the case of lung cancer, diet and body mass index in the case of specific metabolic disorders). However, these are very important elements in determining the health status of patients and their mortality risk (see, for instance, Balia & Jones, 2008; Di Novi, 2010).

Generally, this information appears in detail in surveys conducted with the aim of studying the health of individuals and their access to health services. For example, the American Behavioral Risk Factor Surveillance System Survey (BRFSS), the world's largest cross-sectional telephone survey, conducted every year since 1984 by American state health departments in collaboration with the Centers for Disease Control and Prevention, provides detailed information about health status, diseases, lifestyle and other individual characteristics and can be also combined with other databases, such as environmental information, available from sources like the Environmental Protection Agency, to compare measures of environmental quality and chronic disease. However, despite the BRFSS being one of the most comprehensive surveys in terms of health and lifestyle information, it has significant limitations that are common to all surveys, such as recall bias and inaccurate or low response rates

in the reporting of sensitive information, such as income. Moreover, health indicators use of health services and lifestyles are all self-reported measures and can be subject to measurement errors caused by poorly designed questionnaires, misunderstood concepts, inadequately trained interviewers, differing conceptions of health, varying expectations for one's own health or financial incentives to report ill health (Bago d'Uva et al., 2008; Rossouw et al., 2018). These remain significant limitations for health data that is collected through surveys, thereby making administrative data more attractive from this perspective.

Administrative data remain valuable for policy analysis in healthcare and health research due to the wealth of information that they provide. However, it is crucial to acknowledge another significant concern regarding the sensitivity of administrative health and healthcare data. Rigorous measures are necessary to prevent unauthorised access and to safeguard the privacy of individuals (such as patients or doctors) and entities (such as hospitals). Legal requirements and ethical considerations drive efforts to maintain confidentiality, even in the absence of precise legislation governing data use. These protective measures can lead to delays and can also introduce uncertainty. Accessing administrative health and healthcare data can sometimes be more complicated compared to a survey, particularly for researchers unaffiliated with the data-generating institution, due to privacy concerns and because health data are particularly sensitive (Kapteyn & Ypma, 2007).

One possible way through which to make administrative health data more easily accessible while addressing privacy concerns is by using aggregated administrative data (e.g., averaging by year, gender, geographic area or in other ways) (Cole et al., 2020). A significant example of aggregated data involves population-based cancer registries that gather information on the incidence and survival rates of all reportable tumours within a specific geographic area. Data are sourced from medical records and administrative sources like hospital discharge records and drug prescriptions.

The European Network of Cancer Registries (ENCR) was founded in 1990, following the European Commission's "Europe Against Cancer" Programme in 1985, to enhance data quality and comparability, monitor cancer incidence and mortality and to promote the use of cancer registry data in cancer control and research. The ENCR Secretariat, hosted by the Joint Research Centre (JRC) (in Ispra, Italy) since 2012, supports data dissemination and harmonisation throughout Europe. The JRC also oversees the European Cancer Information System (ECIS), which processes and disseminates standardised cancer data to facilitate accurate comparisons across Europe. This information is easily accessible through an online platform. Specifically, the ECIS web application provides registry-level outputs that include aggregated data on cancer incidence and survival rates and is categorised by cancer type, gender, age group and geographical region (see Giusti et al., 2023). Today, ECIS sets the standard for both monitoring and reporting current statistics on the cancer burden at the European level. It allows for the exploration of geographical and temporal patterns of incidence, mortality and survival across Europe for major cancer types.

However, it is crucial to recognise a significant limitation when using these data for research purposes—whether to evaluate policy interventions or address other

research questions: aggregated data cannot account for individual-specific characteristics such as age, genetics, gender, race or personal variables influenced by lifestyle choices (e.g., smoking, alcohol consumption and environmental factors). Moreover, aggregated data may fail to capture various factors that affect cancer incidence and survival, including social conditions (such as poverty, discrimination and inequality) and access to healthcare services. Additionally, aggregate data do not allow for the identification of treated and control units in most cases and this prevents the estimation of Causal Intervention Effects (CIE).

4.2 Examples of Use

Administrative data have been used extensively in health economics research. This section explores their application through two examples: one concerning primary healthcare access and the other focusing on access to medicines.

Ensuring affordable healthcare access is a primary goal for most European healthcare systems. Primary health care is considered to be the cornerstone of universal health coverage and provides essential and accessible services to individuals and families within their communities. For instance, all of Italy's citizens have access to free primary and secondary care through the National Health Service (NHS), though there may be co-payments for certain prescriptions.

By contrast, in countries without universal health care, such as Ireland, publicly funded programmes help expand healthcare access for lower-income groups or those with high healthcare needs. One such programme is the "medical card," awarded based on an income means test, which allows eligible individuals to receive free public health services. At the beginning of 2000, eligibility for the medical card was extended to people aged 70 and over, regardless of their income. Nolan (2011) used this extension as a natural experiment to study the impact of free General Practitioner (GP) services on avoidable hospitalisations—i.e., hospitalisations potentially preventable with timely and effective primary care or more appropriately treated in a primary care setting. Nolan utilised hospital discharge data from the Irish Hospital In-Patient Enquiry (HIPE), a computer-based system that collects demographic, clinical and administrative data on all day and in-patient discharges (including deaths) from acute hospitals in Ireland, between 1999 and 2004. Nolan tested whether individuals over 70 experienced a significant decrease in avoidable hospitalisations following the medical card eligibility extension. The results indicated that avoidable hospitalisations among those over 70 declined after the extension, as expected as a result of improved access to free GP services. However, this decline was not, on average, significantly different from that experienced by individuals under 70.

Another crucial aspect of enhancing universal health coverage concerns access to affordable and high-quality medicines and health products. Disparities in accessing medicines have often been attributed to drug affordability and to patient-related factors, such as socioeconomic status, health conditions or prescribed therapies (Doshi et al., 2009; Gaynor et al., 2006; Osterberg & Blaschke, 2005). However,

geographic and environmental factors may also impact prescription fulfilment. Di Novi et al. (2020) investigated whether geographic barriers affect access to medicine and patients' ability to fill prescriptions, as well as about the impact of potential non-adherence due to these barriers on patients' health. Their study focused particularly on older individuals with CardioVascular Diseases (CVDs), which are a leading causes of death in most Western countries. They used administrative data from the Liguria Region in Italy, which has the highest proportion of individuals over 65 in Europe. For the study, they created a dataset using four different sources of administrative data: the Liguria Hospital Discharge Records (HDRs) database, the Emergency Departments (ED) registry, the Pharmaceutical Registry and the Regional Population Registry, which contain information about the date of death of Ligurian patients. Their findings revealed that geographic barriers to pharmacies negatively affect patients' adherence to medication. According to their results, non-adherence to pharmacological therapy increases the probability of mortality and leads to higher utilisation of other medical services, such as hospitalisations and emergency department visits.

This study proves the benefits of leveraging various administrative records, including the hospital discharge registry linked via an identification code to data from the regional population, the emergency departments and the pharmaceutical registries. The ability to link these different administrative records, which we will discuss further, represents a significant advancement in addressing research questions—and hence knowledge of what works—that span multiple domains.

4.3 Administrative Data in Health Research and COVID-19 Pandemic

Numerous ad hoc surveys were conducted during the COVID-19 pandemic and several papers utilised surveys when providing evidence about the impact of the COVID-19 pandemic. However, administrative data is routinely collected as part of clinical practice within healthcare systems and have proven invaluable for generating real-world evidence. This evidence has informed regulators, supported policy formation and monitoring and has even shaped service delivery. These data have been instrumental in estimating the number of infections, deaths and mortality rates (see, for instance, De Palo, 2021; Berta et al., 2021).

Administrative data have also been particularly useful in estimating the demand for hospital beds, including intensive care units (ICU) beds and in determining associated staff and equipment requirements (Berta et al., 2020). The pressure on hospital ICUs during the pandemic was significant. Predicting hospital service demand relied heavily on estimating the number of patients requiring hospitalisation and their expected length of stay. Information from hospital discharge records was crucial for predicting bed demand and for improving decision-making and contingency planning during this crisis.

Several countries implemented measures, such as restricting international commercial and trading channels in order to contain the virus. Concerns arose regarding a global drug shortage due to increased demand from patient anxiety and stockpiling. The COVID-19 pandemic, catching many unprepared, triggered worldwide panic buying, particularly for essential goods. The outbreak led to a sharp decrease in outpatient healthcare services, while fears of drug shortages prompted patients to temporarily increase their medication stockpiles. Drug shortages can lead to difficulties in obtaining bioequivalent drugs, higher costs and increased risks of adverse patient outcomes. Studies using administrative data have identified groups at higher risk of health deterioration due to drug shortages and reduced medication adherence. These studies provide essential tools for managing emergencies that impact the social, economic and health aspects of countries grappling with the virus (see Di Novi et al., 2022).

Furthermore, the use of administrative data has facilitated the assessment of individuals' clinical profiles, thereby aiding in the assessment of their risk of coronavirus infection (Orlando et al., 2021) and prioritising them for vaccine distribution (Wright et al., 2022).

5 Conclusions

This chapter explored three key research domains—labour, education and health economics—in which the availability of administrative data is essential for the conducting of robust policy evaluation. Although the discussion treated each field as distinct, the optimal scenario would involve the capacity to interconnect various registries to compile comprehensive data across these areas for individual subjects. Indeed, administrative records can be utilised by exploiting individual identifiers that allow different databases containing individual information such as income, education, occupation, health and healthcare access, etc., to be linked.

This potential integration would significantly enhance the value of the data by creating a comprehensive database, which is particularly beneficial for researchers across various fields. By consolidating data, researchers can more effectively explore cross-field research questions.

An exemplary case of this approach is the administrative data collected by the Centraal Bureau voor de Statistiek (CBS) in The Netherlands (see Chap. 21), which is extensively used in labour, education and health economics literature. The CBS provides access to rich, linked data from hospital admission records, national tax and wealth registry files, education information, mortality and municipality registries containing socio-demographic details (age, gender, migration status, marital status, family size, etc.) about Dutch citizens.

An example of such cross-field approaches using CBS data is offered by Scholte et al. (2015). In this specific case, the authors examined the long-term economic impacts of in utero exposure to the Hunger Winter famine on three outcomes related to

individual economic well-being later in life, namely annual labour earnings, employment status and hospitalisation occurrences, with specific reference to cardiovascular diseases and cancer. They linked the population's longitudinal demographic register with corresponding individual records from the national tax database, which contain labour market information and hospital registration databases; this allowed them to extract information on hospitalisation-related diseases by using the CBS' ICD classification in The Netherlands over the period 1999–2005. The comprehensive nature of these linked administrative datasets allowed the authors to address a research question spanning multiple domains and enabled them to provide a more precise assessment of the relevant causal pathways between the Hunger Winter famine and the outcomes considered.

In addition to the CBS, there are other virtuous examples in the landscape of administrative data, such as Statistics Denmark (Chap. 8), Statistic Finland (Chap. 10) and Statistic Sweden (Chap. 28) to mention only a few. Such data enables collaborative and interdisciplinary research efforts across various fields. The integration of different records, and potentially also with data collected through surveys, allows for a more comprehensive understanding to emerge of complex issues and facilitates informed decision-making in diverse areas of study. Among others, CBS registry data can also be used, albeit for a smaller sample, in conjunction with the Survey of Health, Aging and Retirement in Europe (SHARE), supervised by the Mannheim Research Institute for the Economics of Aging (MEA) (Borsch-Supan, & Jurges, 2005) allowing for the strength of both sources to be combined—surveys and administrative data—overcoming the limitations of both approaches in the process (Börsch-Supan et al., 2013).

References

Adermon, A., Laun, L., Lind, P., Olsson, M., Sauermann, J., & Sjögren, A. (2022). Earnings losses and the role of the welfare state during the COVID-19 pandemic: Evidence from Sweden. *Review of Income and Wealth.*

Bago d'Uva, T., Van Doorslaer, E., Lindeboom, M., & O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics, 17*(3), 351–375.

Balia, S., & Jones, A. M. (2008). Mortality, lifestyle and socio-economic status. *Journal of Health Economics, 27*, 1–26.

Berta, P., Bratti, M., Fiorio, C. V., Pisoni, E., & Verzillo, S. (2021). Administrative border effects in Covid-19 related mortality. *Journal of the Royal Statistical Society Series A: Statistics in Society, 188*(4), 1044–1084.

Berta, P., Paruolo, P., Verzillo, S., & Lovaglio, P. G. (2020). A bivariate prediction approach for adapting the health care system response to the spread of COVID-19. *PLoS ONE, 15*(10), Article e0240150.

Bertheau, A., Acabbi, E. M., Barceló, C., Gulyas, A., Lombardi, S., & Saggio, R. (2023). The unequal consequences of job loss across countries. *American Economic Review: Insights, 5*(3), 393–408.

Blanden, J., Doepke, M., & Stuhler, J. (2023). Educational inequality. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (pp. 405–497). Elsevier.

Börsch-Supan, A., & Jurges, H. (Eds.). (2005). *The survey of health, ageing and retirement in Europe—Methodology*. MEA.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., & Zuber, S. (2013). Data resource profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology, 42*(4), 992–1001.

Buckingham, J. (2019). Reading recovery: A failed investment. *Multilit,* February 2019. https://www.cis.org.au/publication/reading-recovery-a-failed-investment/

Bussink, H., Vervliet, T., & Ter Weel, B. (2022). The short-term effect of the COVID-19 crisis on employment probabilities of labour-market entrants in the Netherlands. *De Economist, 170*(2), 279–303.

Campa, P., Roine, J., & Strömberg, S. (2021). Unequal labour market impacts of covid-19 in Sweden–but not between women and men. *Intereconomics, 56*(5), 264–269.

Card, D., Kluve, J., & Weber, A. (2010). Active labour market policy evaluations: A meta-analysis. *The Economic Journal, 120*(548), F452–F477.

Card, D., Kluve, J., & Weber, A. (2018). What works? A meta analysis of recent active labour market program evaluations. *Journal of the European Economic Association, 16*(3), 894–931.

Carneiro, P., & Heckman, J. J. (2002). The evidence on credit constraints in post-secondary schooling. *Economic Journal, 112*(482), 705–734.

Casarico, A., & Lattanzio, S. (2022). The heterogeneous effects of COVID-19 on labour market flows: Evidence from administrative data. *The Journal of Economic Inequality, 20*(3), 537–558.

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest, 19*(1), 5–51.

Chall, J. S. (1967). *Learning to read: The great debate*. McGraw Hill.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2633–2679.

Cole, S., Dhaliwal, I., Sautmann, A., & Vilhuber, L. (2020). *Handbook on using administrative data for research and evidence-based policy*. https://admindatahandbook.mit.edu/book/v1.0-rc5/index.html

Cook, P., Rodes, D. R., & Lipsitz, K. L. (2017). The reading wars and Reading Recovery: What educators, families, and taxpayers should know. *Learning Disabilities: A Multi-Disciplinary Journal, 17*(22), 12–23. https://js.sagamorepub.com/ldmj/article/view/8391

Crato, N. (2021). Setting up the scene: Lessons learned from PISA 2018 statistics and other international student assessments. In N. Crato (Ed.), *Improving a country's education.* Springer.

Crato, N. (2023). From lack of data to data unlocking. In E. Bertoni, M. Fontana, L. Gabrielli, S. Signorelli, & M. Vespe (Eds.), *Handbook of computational social science for policy*. Springer. https://doi.org/10.1007/978-3-031-16624-2_6

Crato, N., & Patrinos, H. A. (2025). *Improving national education systems after Covid 19*. Springer. https://doi.org/10.1007/978-3-031-69284-0

Cuntz, A., & Sahli, M. (2022). *The COVID-19 impact on artistic income: Evidence from Germany. Resilience and Ingenuity Global Innovation Responses to COVID-19.*

Daniel, Koretz. (2008). *Measuring up what educational testing really tells us*. Harvard University Press.

Depalo, D. (2021). True COVID-19 mortality rates from administrative data. *Journal of Population Economics, 34*(1), 253–274.

Di Novi, C. (2010). The influence of traffic-related pollution on individuals' life-style: Results from the BRFSS. *Health Economics, 19*, 1318–1344.

Di Novi, C., Leporatti, L., & Montefiori, M. (2020). Older patients and geographic barriers to pharmacy access: When nonadherence translates to an increased use of other components of health care. *Health Economics, 29*, 97–109.

Di Novi, C., Leporatti, L., Levaggi, R., & Montefiori, M. (2022). Adherence during COVID-19: The role of aging and socio-economics status in shaping drug utilization. *Journal of Economic Behavior & Organization, 204*, 1–14.

Doshi, J. A., Zhu, J., Lee, B. Y., Kimmel, S., & Volpp, K. G. (2009). Impact of a prescription copayment increase on lipid-lowering medication adherence in veterans. *Circulation, 119*(3), 390–397.

Ecalle, J., Gomes, C., Auphan, P., Cros, L., & Magnan, A. (2019). Effects of policy and educational interventions intended to reduce difficulties in literacy skills in grade 1. *Studies in Educational Evaluation, 61*, 12–20.

Engzell, P., Frey, A., & Verhagen, M. D. (2022). *Learning loss due to school closures during the COVID-19 pandemic, PNAS 2021-118*.

European Commission, Directorate-General for Employment, Social Affairs and Inclusion, Pompili, M., Kluve, J., Jessen, J. et al., Meta-analysis of the ESF counterfactual impact evaluations – Final report, Publications Office of the European Union, 2023.

Gaynor, M., Li, J., & Vogt, W. B. (2006). *Is drug coverage a free lunch? Cross-price elasticities and the design of prescription drug benefits* (NBER Working Paper No. 12758).

Giusti, F., Martos, C., Trama, A., Bettio, M., Sanvisens, A., Audisio, R., Arndt, V., Francisci, S., Dochez, C., Ribes, J., Fernández, L. P., Gavin, A., Gatta, G., Marcos-Gragera, R., Lievens, Y., Allemani, C., De Angelis, R., Visser, O., Van Eycken, L., & ENCR Working Group on Treatment Data Harmonisation. (2023). Cancer treatment data available in European cancer registries: Where are we and where are we going? *Frontiers in Oncology, 13*, 1109978.

Gomendio, M. (2021). Spain: The evidence provided by international large-scale assessments about the Spanish education system: Why nobody listens despite all the noise. In: Crato, N. (Ed.), *Improving a country's education*. Springer.

Gomendio, M. (2025). Spain: The response of the education system to the COVID-19 pandemic: how LNOB (Leaving No-One Behind) got lost in translation. In: Crato, N., Patrinos, H. A. (Eds.), *Improving National Education Systems After COVID-19. Evaluating Education: Normative Systems and Institutional Practices*. Springer, Cham. https://doi.org/10.1007/978-3-031-69284-0_12

Gulyas, A., & Pytka, K. (2020). The consequences of the COVID-19 job losses: Who will suffer most and by how much. *Covid Economics, 1*(47), 70–107.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*, 466–479.

Hanushek, E. A., & Woessmann, L. (2020). *The economic impacts of learning losses*. OECD (September).

Hanushek, E.A., Strauss, B. (2025). United States: The size and variation of the pandemic learning losses. In: Crato, N., Patrinos, H. A. (Eds.), *Improving national education systems after COVID-19. Evaluating Education: Normative systems and institutional practices*. Springer. https://doi.org/10.1007/978-3-031-69284-0_13

Hassink, W., Kalb, G. R., & Meekes, J. (2020). *The Dutch labour market early on in the COVID-19 outbreak: Regional coronavirus hotspots and the national lockdown* (Life Course Centre Working Paper (2020–22)).

Heckman, J. J., Lalonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labour market programs. In O. Ashenfelter & D. Card (Eds.), *Handbook of labour economics* (Vol. 3A, pp. 1865–2095). Elsevier.

Hensvik, L., Le Barbanchon, T., & Rathelot, R. (2021). Job search during the COVID-19 crisis. *Journal of Public Economics, 194*, Article 104349.

Herrera-Pavo, M.Á., Jaramillo-Baquerizo, C., Valencia, V.H. (2025). Ecuador: Inserting the country into the PISA experience. In: Crato, N., Patrinos, H. A. (Eds.), *Improving national education systems after COVID-19. Evaluating education: Normative systems and institutional practices*. Springer. https://doi.org/10.1007/978-3-031-69284-0_3.

Jakubowski, M., Gajderowicz, T., & Patrinos, H. A. (2023). Global learning loss in student achievement: First estimates using comparable reading scores. *Economics Letters, 232*, Article 111313.

Jakubowski, M., Gajderowicz, T., & Patrinos, H. A. (2024). *COVID-19, school closures, and student learning outcomes: New global evidence from PISA* (IZA Discussion Paper No. 16731).

Kapteyn, A., & Ypma, J. Y. (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labour Economics, 25*(3), 513–551.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180–195.

Lane, H. B., Contesse, V. A., Gage, N. A., & Burns, M. K. (2025). Effect of an instructional program in foundational reading skills on early literacy development of students in kindergarten and first grade. *Reading Research Quarterly, 60*, Article e607. https://doi.org/10.1002/rrq.607

Loosveldt, G., & Storms, V. (2008). Measuring public opinions about surveys. *International Journal of Public Opinion Research, 20*(1), 74–89.

Maldonado, J. E., & De Witte, K. (2022). The effect of school closures on standardised student test outcomes. *British Educational Research Journal, 48*, 49–94.

Marôco, J. (2025). Portugal: COVID-19 learning losses through the lenses of ILSA and local low stake assessments. In: Crato, N., Patrinos, H. A. (Eds.), *Improving national education systems after COVID-19. Evaluating education: Normative systems and institutional practices.* Springer. https://doi.org/10.1007/978-3-031-69284-0_10

Moore, J. C., Stinson, L. L., & Welniak, E. J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics-Stockholm, 16*(4), 331–362.

Nolan, A. (2011). An extension in eligibility for free primary care and avoidable hospitalisations: A natural experiment. *Social Science & Medicine, 73*(7), 978–985.

Nunes, C., Carvalho, B. P., Pereira dos Santos, J., Peralta, S., & Tavares, J. (2023). Failing young and temporary workers? The impact of a disruptive crisis on a dual labour market. *The BE Journal of Economic Analysis & Policy, 23*(2), 349–395.

Orlando, V., Rea, F., Savaré, L., Guarino, I., Mucherino, S., Perrella, A., Trama, U., Coscioni, E., Menditto, E., & Corrao, G. (2021). Development and validation of a clinical risk score to predict the risk of SARS-CoV-2 infection from administrative data: A population-based cohort study from Italy. *PLoS ONE, 16*(1), Article e0237202.

Osterberg, L., & Blaschke, T. (2005). Adherence to medication. *The New England Journal of Medicine, 353*, 487–497.

Quan, H., Parsons, G. A., & Ghali, W. A. (2004). Validity of procedure codes in International Classification of Diseases, 9th revision, clinical modification administrative data. *Medical Care, 42*(8), 801–809.

Rossouw, L., Bago d'Uva, T., & Van Doorslaer, E. (2018). Poor health reporting? Using anchoring vignettes to uncover health disparities by wealth and race. *Demography, 55*(5), 1935–1956.

Sánchez, J. P., Koskinen, J., & Plewis, I. (2014). Measurement error in retrospective work histories. In *Survey research methods* (Vol. 8, No. 1, pp. 43–55).

Scholte, R. S., Van Den Berg, G. J., & Lindeboom, M. (2015). Long-run effects of gestation during the Dutch Hunger Winter famine on labour market and hospitalization outcomes. *Journal of Health Economics, 39*, 17–30.

Tire, G. (2025). Estonia: The Impact of Covid-19 on Education in Estonia—What Have We Learned from PISA 2022 and National External Evaluations? In: Crato, N., Patrinos, H. A. (Eds.), *Improving national education systems after COVID-19. Evaluating education: Normative systems and institutional practices.* Springer. https://doi.org/10.1007/978-3-031-69284-0_5

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859.

Wright, B. J., Tideman, S., Diaz, G. A., French, T., Parsons, G. T., & Robicsek, A. (2022). Comparative vaccine effectiveness against severe COVID-19 over time in US hospital administrative data: A case-control study. *The Lancet Respiratory Medicine, 10*(6), 557–565.

Yan, T., & Curtin, R. (2010). The relation between unit nonresponse and item nonresponse: A response continuum perspective. *International Journal of Public Opinion Research, 22*(4), 535–551.

Nuno Crato CGIP, GCHP, is Research Professor of Mathematics and Statistics, Cemapre, ISEG, University of Lisbon and member of the French government Scientific Council for the National Education (CSEN). Works on stochastic models, time series applications on financial and social problems, and statistical data-based evaluation of policy measures, namely education. He has published extensively on econometrics and statistics and worked on education policy analysis. President of the Portuguese Mathematical Society (2004–2010) and director of the International Institute of Forecasters (2016–2020), is founding organizer of LESE, the Lisbon Economics and Statistics of Education. From 2011 to 2015, he was Education and Science Minister of Portugal.

Cinzia Di Novi is an Associate Professor of Public Economics at the Department of Economics and Management, University of Pavia, and the current President of the Italian Health Economics Association (AIES) for the 2024–2026 term. From 2021 to 2024, she served as a Quantitative Policy Analyst on Impact Evaluation at the Joint Research Centre of the European Commission (JRC CC-ME) in Ispra. She is also a member of the Executive Committee of the Society of Economics of the Household (SEHO) and an External Member of the Health, Econometrics, and Data Group (HEDG) at the University of York, as well as a member of the Scientific Committee of the APHEC Research Centre at the University of Genoa. Her work is characterized by concrete experience in Health Impact Assessment (HIA) with a strong focus on integrating health considerations into policymaking. Her research focuses on health economics, the economics of aging, policy evaluation, and applied micro-econometrics.

Elena Claudia Meroni is a Senior Scientific Policy Analyst at the Competence Centre on Microeconomic Evaluation (CC-ME) of the Joint Research Centre of the European Commission. Her research focuses on evaluating public policies in the fields of education and the labour market. Elena has a PhD in Statistics from the University of Padua and a Master degree in Economics from Bocconi University.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 31
Learning About Environmental Outcomes in Europe



Zelda Brutti, Marzia Freo, and Laura Serlenga

Abstract The access to high-quality, interoperable and granular data is key to the advancement of research and policy analysis in the environmental domain, which in turn aids steering Europe towards fulfilment of Green Deal objectives. An adaptive EU legal framework and the combined efforts of Member States and other data-holding institutions are supporting the creation and improvement of harmonised environmental data. Several hurdles persist, including (i) heterogeneity in data harmonisation and sharing practices across countries; (ii) delays between data collection and data availability for analysis; (iii) the challenging integration of environmental data stemming from mixed sources. Every improvement along these dimensions appears a worthy undertaking, given the superiority of insights obtained from analyses based on high-quality data incorporating administrative environmental sources.

Z. Brutti · M. Freo (✉) · L. Serlenga
European Commission Joint Research Centre, Ispra, Italy
e-mail: marzia.freo@unibo.it

Z. Brutti
e-mail: zelda.brutti@gmail.com

L. Serlenga
e-mail: laura.serlenga@uniba.it

© The European Union 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_31

1 Introduction

This chapter discusses the potential for better use of data and analysis for learning what works to aid steering Europe towards environmental objectives, as set out in the European Green Deal. These objectives are important, both in financial terms[1] and in the fight against climate change.[2]

In line with the Global Green New Deal and the European Green Deal policy objectives, European Countries have been increasing environmental protection expenditure (EPE), which encompasses ambient air, soil and water; wastewater and waste management; noise abatement; biodiversity protection; protection against radiation; and environmental research and development (R&D). Other types of green expenditure such as on renewables, energy and resource efficiency, and the circular economy transition have also been growing.

The EU's 2021–2027 budget and the 2021–2026 EU Recovery and Resilience Facility (RRF) have earmarked additional funding for climate- and biodiversity-related activities, so that the increasing trend of EPE is likely to continue over the coming years (European Environment Agency, 2023). These spending dynamics are only examples of a globally heightened awareness of the need to address environmental challenges and protect the planet for future generations.

The next section of this chapter highlights a number of developments in the EU legal framework concerning data sharing, data interoperability, and public access to information, especially for what geospatial datasets are concerned. The recent introduction of the INSPIRE data sharing portal and the conceptualisation of High Value Datasets are examples of such advancements and represent elements that are likely to boost the availability of key data sources for environmental research and analysis, thus improving the quality of such endeavours. Several challenges remain in achieving the goal of a unified, standardised, and openly accessible EU-wide dataset, which provides comprehensive environmental information. On the one hand, there are technical challenges such as the definition of data harmonisation protocols able to guarantee interoperability and comparability of environmental data across the EU, as well as the integration of information across different environmental domains. On the other hand, there are country-specific issues, such as the speed at which national and subnational data infrastructures are able to adopt EU data standards and protocols, and eventually create internationally sharable data products. Finally, there are challenges related to the nature itself of environmental data, which often involves lengthy processes of sampling, collection, laboratory analysis and testing, before being ready for publication in research databases. Coupled with the

[1] As of 2020, to achieve the goals set by the European Green Deal, the Commission has pledged to mobilise at least €1 trillion in sustainable investments over the next decade. Thirty percent of the EU's multiannual budget (2021–2028) and the EU's unique NextGenerationEU (NGEU) instrument to recover from the COVID-19 pandemic, has been allocated for green investments (European Commission, 2020).

[2] The Intergovernmental Panel on Climate Change (IPCC) of the United Nations finds that there are positive signs and increased evidence of climate action and that some countries have already achieved a steady decrease in emissions consistent with limiting warming to 2 °C (IPCC, 2022).

fact that environmental changes tend to happen slowly over time and feature cumulative patterns, the lengthy data preparation results in large lags between human action and a quantification of its impacts on the environment.

A particularly valuable subgroup of environmental datasets are those from *administrative* information sources, i.e. data collected in a systematic and standardised way, with broad or even universal geographical coverage, often for purposes that go beyond environmental research. Information extracted from these sets tends to feature desirable properties such as precision and less reporting bias with respect to survey sources, and standard and over-time comparable collection or measurement protocols. Be it for the purpose of descriptive analysis, for predictive modelling or for ex-post impact evaluation, universal, and granular administrative data is able to deliver policy-relevant insights that would otherwise go undetected, as exemplified in a short case study in Sect. 3.

Finally, special emphasis should be placed on spatial environmental datasets, as they offer valuable georeferenced environmental information. Being able to account for exact spatial location is crucial when working with environmental data, as environmental processes are spatially interconnected, featuring complex interactions among various natural components and policy interventions. As the availability of spatial datasets increases, a big challenge is integrating data from different sources with varying spatial resolutions. Section 4 discusses the issue in further detail, reflecting on the advantages and drawbacks of the most commonly adopted solutions.

The rest of the chapter is organised as follows: Sect. 2 provides an overview of recent developments and persisting challenges in EU environmental data availability; Sect. 3 discusses the main avenues of application of high-quality environmental data and provides a practical example of the value of high-quality administrative information; Sect. 4 discusses challenges and solutions to spatial data integration. Section 5 concludes.

2 Policy Context

The concept of 'environmental information' was laid down by the 1998 Aarhus Convention and has since then evolved in the EU legal framework, gradually expanding from data describing the state of the environment (such as atmosphere, water, soil, biodiversity) to also include data about factors that are 'likely' to affect the environment even indirectly, such as greenhouse gas emissions or noise and waste.

At EU level, there are a number of longer-running legal frameworks that support the harnessing and access to administrative public sector data, such as the Freedom of Access to Information Directive, the INSPIRE Directive and the Open Data Directive, all of which pursue the goal of establishing common data infrastructures.[3] More

[3] Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information and repealing Council Directive 90/313/EEC [2003] OJ L041 ('Freedom of Access to Information Directive'). Directive 2007/2/EC of the European Parliament

recent and more focused on the environmental domain is the Green Deal Data Space (GDDS); this is one of the common European data spaces envisioned by the European Commission to integrate large amounts of cross-sectorial data in support of the priority actions of the Green Deal. The purpose of this initiative is to connect environmental data from existing but fragmented ecosystems and dataspaces from both private and public sources across Member States, and feed it into new services and applications that contribute to the objectives of the Green Deal, by enabling its access, sharing, and re-use for research and policy.

One of the most attractive outcomes of the recent developments on the EU data scene is the establishment of a list of so-called 'High Value Datasets', through the adoption of the High Value Datasets Implementing Act by the European Commission in December 2022, as a follow-up implementing decisions to the Open Data Directive (European Commission, 2022a). High-value datasets are defined based on their potential to provide essential benefits to society, the environment and the economy, so that the public sector bodies holding them are required to guarantee specific conditions in terms of accessibility for re-use and minimum requirements for data dissemination in machine-readable format and via application programming interfaces (APIs). High Value Datasets are divided into six thematic categories: *Statistics*, *Companies and company ownership*, *Meteorological*, *Geospatial*, *Earth observation and environment* and *Mobility*, several of which have a strong potential to support research in the environmental domain. In particular, the latter three categories include datasets within the scope of the INSPIRE data themes, which can play an essential role as a public sector contribution to the Green Deal Data Space. INSPIRE (Infrastructure for Spatial Information in Europe) collects the *spatial* data infrastructure that is specifically designed for the purposes of EU environmental policies and policies or activities which may have an impact on the environment. Table 31.1 provides an overview of the High-Value Datasets under the scope of INSPIRE, dividing them into three thematic categories and highlighting those datasets that are particularly preeminent from a policymaking perspective in the left panel, where applicable. The 2022 evaluation of INSPIRE ordered by the Commission (European Commission, 2022b) has provided an overall positive picture of the current situation, finding that the implementation of the INSPIRE Directive is continuously improving, that the initiative has led to an increased availability and better access to spatial data and services across the EU and that its objectives remain highly relevant. On the other hand, the evaluation found that the INSPIRE implementation can be improved along a number of dimensions, such as the level of access of the spatial data for viewing and downloading (with around half of the documented datasets not yet accessible), the conformity in metadata specifications and in data harmonisation; in all of these aspects, there is considerable heterogeneity in the progress achieved across different Member States. The study also highlighted the fact that some Member States have

and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community [2007] OJ L108 ('INSPIRE'). Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast) PE/28/2019/REV/1 [2019] OJ L172 ('Open Data Directive').

developed two different spatial data infrastructures, one for national use and one for the implementation of the INSPIRE Directive, which would appear as an unnecessary duplication of data creation and maintenance costs. Finally, there are significant discrepancies in the number of datasets across Member States, their geographical scope (national, regional, local) and their content, yielding a current geographical coverage situation that is judged as 'not yet optimal'.

In the final part of this introduction, it may be worth adding some additional reflections on data-related challenges when it comes to empirical research in the environmental domain. The perhaps most attractive but also most demanding research exercise consists in identifying a causal chain running from environmental policies, through behavioural responses by individual units of observation (firms, farms, individual citizens, etc.) and culminating in an impact on environmental outcomes. Being able to pin down such causal link empirically requires being able to access data on

Table 31.1 High-value datasets under the scope of INSPIRE

Earth observation and environment	
Priority Datasets for eReporting	*Other High-value datasets*
• Air • Nature • Water • Noise • Waste • Industrial emissions • Industrial accidents	• Hydrography • Protected sites • Elevation • Geology • Land Cover • Orthoimagery • Area management • Bio-geographical regions • Environmental monitoring facilities • Energy resources • Habitats and biotopes • Land use • Mineral resources • Natural risk zones • Oceanographic features • Production and industrial facilities • Species distribution • Soil • Sea regions
Geospatial	
Agricultural parcels	*Other Core reference datasets*
• GSAA (Geospatial Aid Application) • LPIS (Land Parcel Identification System)	• Addresses • Administrative units • Cadastral parcels • Geographical names • Buildings
Mobility	
• Transport Networks	

Source European Data Portal, European Commission

all the three stages of the chain—the policy implementation stage, the behavioural response stage and the environmental outcomes stage—and, most importantly, being able to link them to one another. Linking may be performed at the individual-unit level—e.g. an environmental policy applied to individual firms getting linked to each firm's behaviours and then to environmental measures taken on those same firms—or at the geographical level—e.g. an environmental policy applied to specific local areas getting linked to environmentally relevant behaviours and to environmental outcomes in that area. The general principle that ought to be followed is 'the more micro-level the link, the better for evaluation purpose', since operating with data aggregated and averaged into larger clusters often goes hand in hand with loss of precision and of information on impact heterogeneity. There is a continuously expanding array of environmental data available for research, coupled with precise geo-location or individual identifiers. However, there is still large heterogeneity in the progress towards the full implementation of the objectives laid out by INSPIRE and the other EU-level data sharing frameworks, a fact that often induces the analyst to focus on case studies with a narrow geographical scope, in exchange for the benefit of working with high-quality data.

Among data-related challenges, it is also worth reminding that the timely availability of data is an indispensable condition for maintaining the relevance of empirical research findings, especially when such findings are meant to provide evidence-based advice to future policy programming periods. In the environmental domain, the process from data collection to availability of the final data product for research can be lengthy, often spanning several years, and thus represents an obstacle to producing timely research results. To give a concrete example, at the time of writing the LUCAS Soil Modules, which provide highly valuable soil quality measurements and are managed by the European Soil Data Centre (ESDAC), have been released in three waves (2009/12, 2015 and 2018). The 2018 wave of soil sampling was followed by laboratory analyses, data processing and data release procedures that resulted in the data being available for researchers in 2022, that is, four years later; the 2022 wave is still being processed. This example just illustrated the point that the time gap between the implementation of environmental policies and their measurable consequences is oftentimes a critical consideration in environmental research and this consideration emphasises the need for patience, long-term commitment, and robust environmental monitoring programs and data-sharing platforms.

To conclude, the current dynamic of European-level environmental data availability has been displaying a clear positive trend for several years now. The amount of precise, timely, and sophisticated environmental outcome data on a European scale is constantly increasing, thanks to the establishment of foresighted legal frameworks and the combined efforts of Member States as well as institutions such as the European Environmental Agency, the Joint Research Centre, and the European Space Agency, just to mention a few. It is imperative that this positive trend in data availability is not only maintained but also accelerated; the wealth of data generated by public and private institutions alike needs to be harnessed to provide policymakers, researchers, and stakeholders with the necessary tools to comprehensively monitor the EU environment as well as evaluating the impact of environmentally targeted

policies. The number of opportunities for high-quality research, robust assessments and evaluations, and cost-effective policymaking is boosted immensely by the availability of frameworks for distributing environmentally relevant data according to FAIR (Findable, Accessible, Interoperable, Reusable) principles.

The next section provides further discussion targeted at highlighting the specific advantages of environmentally relevant data of *administrative* type, as compared to survey-based or other less systematic sources, followed by an empirical illustration based on the agricultural domain.

3 The Value of Data for Environmental Policy Analysis

Access to high-quality and interoperable data represents the backbone of the capacity expansion of environmental research and development and the support of evidence-based policymaking (Finck & Mueller, 2023). The wider the range of data sources available to the analyst, the higher the chance of gaining multidimensional insights into the topic of interest and of uncovering original evidence to support decision-making processes.

Through **descriptive analyses, monitoring exercises, and data visualisation tools**, environmental data is summarised and presented in a way that helps identifying patterns, trends and relationships; understanding environmental changes over time and space; detecting any anomalies which may require further investigation or action. Figure 31.1 represents an example of retrospective monitoring exercise: the figure summarises the changes in water pollutants release from 2010 to 2022 in the different EU Member States and uses colour labels to enable a rapid appraisal by the reading audience.

Predictive analyses take descriptive analyses further and are used to forecast future trends or scenarios based on historical and current environmental data. Various statistical and machine learning techniques can be employed to project identified patterns and relationships into the future, to make informed decisions and develop strategies for managing and protecting the environment. Figure 31.2 summarises the results of a complex forecasting exercise on the days of high-to-extreme fire danger one should expect over the coming decades under three different climate change scenarios (+1.5 °C, +2.0 °C, +3.0 °C), and informs on the geographical heterogeneity of the risk across the European Union. The data represented in the figure represents median ensemble of five different climate models, which are coupled with simulated patterns of temperature and precipitation and several other factors affecting fire danger.

Ex-post and causal evaluation analyses in the environmental domain are used to assess the effectiveness, efficiency, and impact of environmental policies, programs, and interventions after they have been implemented. Such evaluations help to determine the cause-and-effect relationships between actions taken and observed outcomes, enabling decision-makers to learn from past experiences and improve future initiatives, ensuring the efficient and effective use of resources in

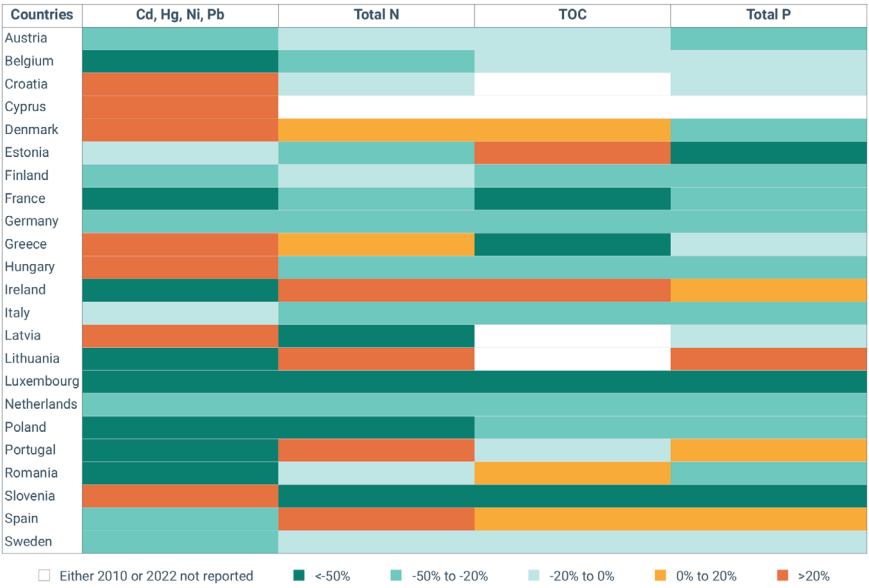


Fig. 31.1 Water pollutant releases changes from 2010 to 2022 for the EU Member States. *Note* TOC—total organic carbon, Total N—total nitrogen, Total P—total phosphorous, Cd—Cadmium, Hg—Mercury, Ni—Nickel, Pb—Lead. Croatia comparison is from 2014 because it was not an EU member state before. *Source* European Environment Agency (Published 31 May 2024)

addressing environmental challenges. Figure 31.3 illustrates an example of counterfactual impact evaluation exercise, in which the effectiveness of a Swedish carbon tax policy that was introduced in the early 1990s is quantified, using panel data on past emissions and quasi-experimental evaluation methodologies.

Administrative environmental data is collected by government agencies or other administrative bodies as part of routine operations and functions related to environmental management and monitoring, in a systematic and recurrent manner. As such, there is a range of benefits specific to this type of data, or at least not commonly offered by alternative information sources such as survey exercises. Many of such benefits in fact characterise administrative data *in general*, whichever policy topic one may be considering; however, the following paragraphs will be dedicated to a brief review of those dimensions that are particularly salient when operating in the environmental domain. Following such recap, the second part of this section provides a practical application of the aspects discussed, through the illustration of a policy analysis case study that highlights the difference in insights an analyst can obtain, depending on the data used for the study.

Coverage, granularity and accuracy of information are dimensions in which administrative environmental data, acquired through a well-established and comprehensive reporting network, tends to outperform the output of most survey-based sources. Under the coverage and granularity aspect, administrative records are less

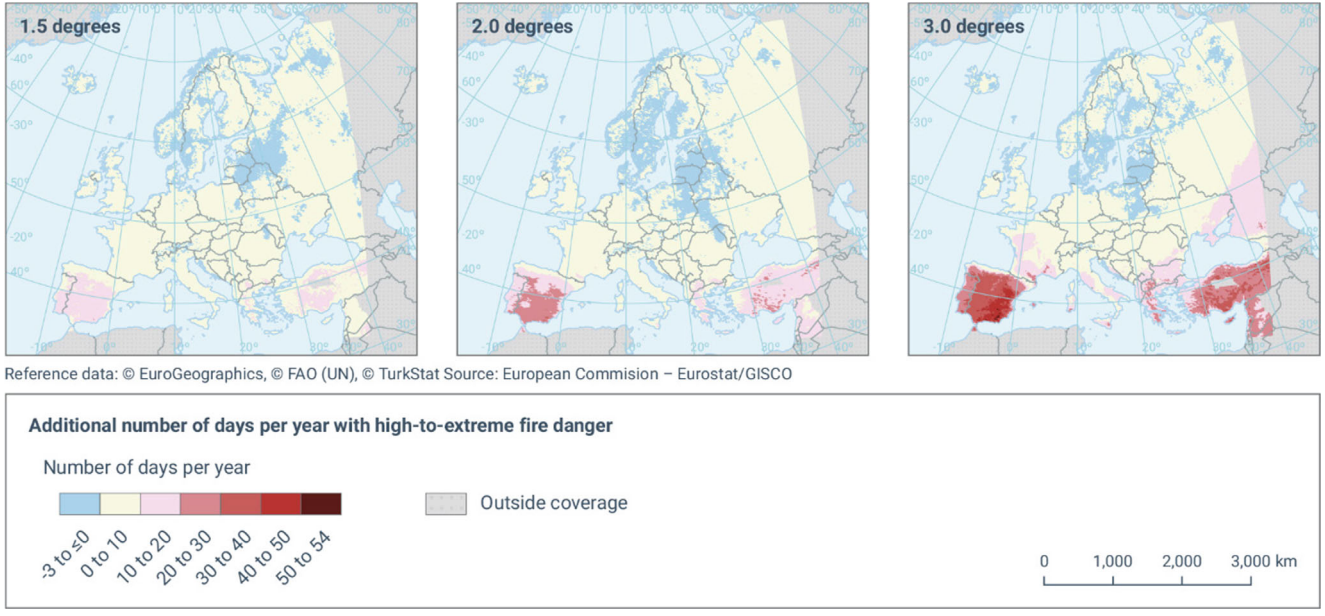


Fig. 31.2 Additional number of days per year with high-to-extreme fire danger. *Note* Additional number of days per year with high-to-extreme fire danger by weather (daily Fire Weather Index ≥30) for different levels of global warming compared to the reference time (1981–2010). Median ensemble of five model-statistics. The maximum extent of time ranges for each global-warming level are as follows: +1.5 °C: 2007–2044; +2.0 °C: 2016–2059; +3.0 °C: 2037–2082. See reference publications: de Rigo et al. (2017) and Costa et al. (2020). *Source* European Environment Agency (Published 01 Jul 2024)

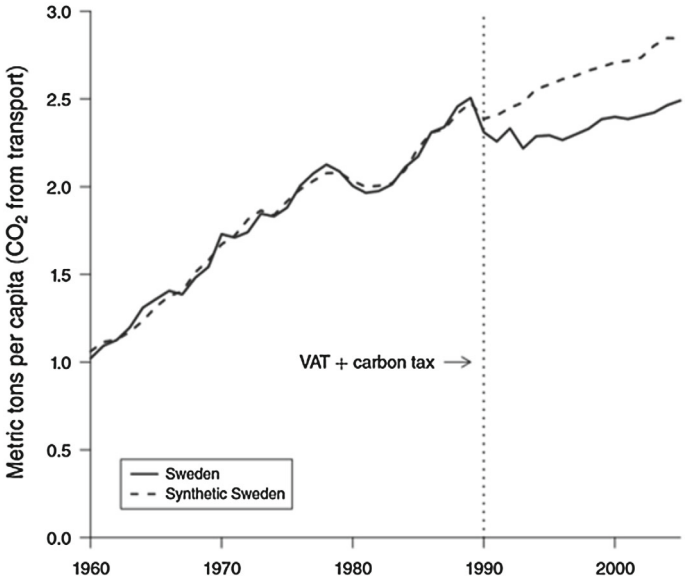


Fig. 31.3 The impact of carbon taxes on CO2 emissions in Sweden. *Note* The figure illustrates the effect of the implementation of a carbon tax and a value-added tax on transport fuel on carbon dioxide emissions from transport in Sweden. The counterfactual no-policy scenario ("Synthetic Sweden") is constructed from a comparable group of OECD countries. *Source* Andersson (2019)

likely to suffer from *selection and representativeness issues*, which typically characterise information sourced through survey samples. Administrative data often covers the entire population of interest—such as land feature databases derived from the Copernicus Land Monitoring Services or other satellite imagery programmes, which routinely output geographical information on land-cover and its changes on a global scale and at a highly detailed level. In other cases, administrative data coverage is not universal but nevertheless designed according to criteria that ensure minimum representativeness standards as well as comparability across time and geographical units—an example of which is represented by the air quality assessment data reported by Member States to the European Commission under the Ambient Air Quality Directives.[4]

Now turning to the accuracy aspect, one first consideration to make is that the *objectivity* of administrative environmental data is backed by the fact that these are collected by organisations which are essentially responsible for monitoring and maintaining records, rather than having any personal interest in the outcomes reported. On the contrary, environmental survey respondents such as businesses, farmers or

[4] The Ambient Air Quality Directives (Directive 2008/50/EC and Directive 2004/107/EC) set standards for ambient air quality for 13 air pollutants to be attained by all Member States across their territories against specified timelines and common methods and criteria to assess air quality in all Member States in a comparable and reliable manner. The Directives also include provisions for periodical results reporting and corrective action plans when needed.

landowners may be influenced by different kinds of reporting bias. Moreover, respondent's perception, understanding, and interpretation of questions are also elements that may induce accuracy losses that are generally avoided in the centralised and standardised collection procedures that characterise administrative data.

A further important dimension to consider is *timeliness of the information*, which in many cases administrative data sources can guarantee through real-time or near-real time reporting systems, as opposed to the usually lengthy administration and data processing times required by survey exercises. Finally, *cost-effectiveness* may be an advantage when administrative data is used for environmental policy assessment or evaluation exercises, as the data is already available and maintained by monitoring organisations for other purposes, whereas environmental surveys require additional resources, such as designing questionnaires, sampling, data collection, and ad-hoc data analysis.

In summary, the specific value of being able to leverage administrative data when analysing environmental policy outcomes lies in its coverage and granularity, accuracy and objectivity, timeliness, and cost-effectiveness. These advantages can help researchers obtain more comprehensive insights into the current state of environmental features of interest, as well as into the effects of environmental policies, thus enabling policymakers to make more informed decisions.

3.1 An Illustration on Environmental Policy Analysis

As an illustration of the value of universal coverage, granularity and accuracy of administrative data in environmental applications, this section briefly presents a real-world environmental policy evaluation setting and compares the picture obtained using data from a large, representative survey to the one obtained leveraging universal administrative data.

The Common Agricultural Policy (CAP) programming period that started in 2014 and ran until 2022 had stepped up the ambition in terms of environmental goals and introduced a so-called 'Greening' package, which accounted for approximately 30% of direct subsidies to EU farmers over those years. The Greening policy package was composed of a number of environmental obligations farmers had to comply with in order to qualify for subsidies and it has been far from uncontroversial, on the one hand perceived as burdensome in terms of implementation by farmers and in terms of monitoring by national authorities, and on the other hand deemed not ambitious enough by several environmental associations and interest groups, auditing bodies, and academic research. As a consequence, there has been a strong interest by the European Commission and by several other institutions to understand whether the environmentally-friendly practices composing this policy package have been producing the desired outcomes in terms of safeguarding soil quality and biodiversity.

Investigating the effects of Greening is far from straightforward due to the complex nature of the environmental subject matter and, most importantly, due to the scarcity of data available to analysts and researchers. The ideal set of data needed for ex-post

evaluation of environmental impacts of Greening is one that links information on the actual implementation of Greening provisions to observational environmental outcomes at the very local level, ideally at the individual farm level, with a wide-enough temporal coverage. An overview of environmental data sources available has already been discussed in the first section of this chapter and will not be the main object of focus here. In terms of data on EU farms, the two sources that are most popular among researchers are the FADN (Farm Accountancy Data Network) and the FSS (Farm Structure Survey), both of which share the feature of being large and rigorous survey exercises carried out throughout the EU with a common methodology and on a regular basis, which provide statistics on farms and farming activities that are representative of the full population and comparable across countries and time.

For the purpose of this section, we have extracted data on Spanish farms recorded in the 2021 FADN survey and assume we are taking the first steps towards an environmental impact evaluation of Greening obligations in this country. It is sufficient to know that the two most salient obligations apply to farms owning more than 10 hectares (the *crop diversification* requirement) and more than 15 hectares (the *ecological focus area* requirement) of arable land, so that we focus our attention on farms around that size threshold; the researcher may plan to compare environmental outcomes measured on farms that were obliged to comply with Greening with those that did not, to obtain insights on the effectiveness of the policy. Figure 31.4 shows the distribution of Spanish farm holdings that are included in the 2021 FADN survey and whose declared arable land size lies between 5 and 20 hectares, i.e. approximately 1,470 firms. In the figure, farms are grouped in size bins of half a hectare and the dashed vertical lines mark the size thresholds at which Greening obligations are triggered.

By looking at this figure, the analyst may notice an up-and-down pattern that suggests the presence of round-number bias in arable land size reporting, by which the size bins that include round sizes (5 to 5.4, 6 to 6.4, 7 to 7.4… hectares) are systematically higher than their complements (5.5 to 5.9, 6.5 to 6.9, 7.5 to 7.9… hectares). Rounding figures when reporting them to answer questionnaires is a quite common phenomenon which is part of the larger class of *reporting biases* affecting survey data, and it may cause difficulties for impact evaluation designs, especially if the policy under examination is based on a precise threshold like the two Greening obligations we are considering in this illustration.[5]

Apart from the observation of a probable tendency by farmers or surveyors to round the holding's arable land size when filling out the FADN questionnaire, no other salient features emerge from the picture. In particular, there is no evidence of any policy avoidance behaviour in the farm population, as the number of farms 'just below' 10 or 15 hectares is not unusually high, even though the individual incentives for manipulating one's arable land size in order to remain below these thresholds

[5] Impact evaluation literature has been developing solutions to overcome the problem of rounded data in empirical applications. For example, see Dong, Y. (2015). Regression discontinuity applications with rounding errors in the running variable. *Journal of Applied Econometrics, 30*(3), 422–446.

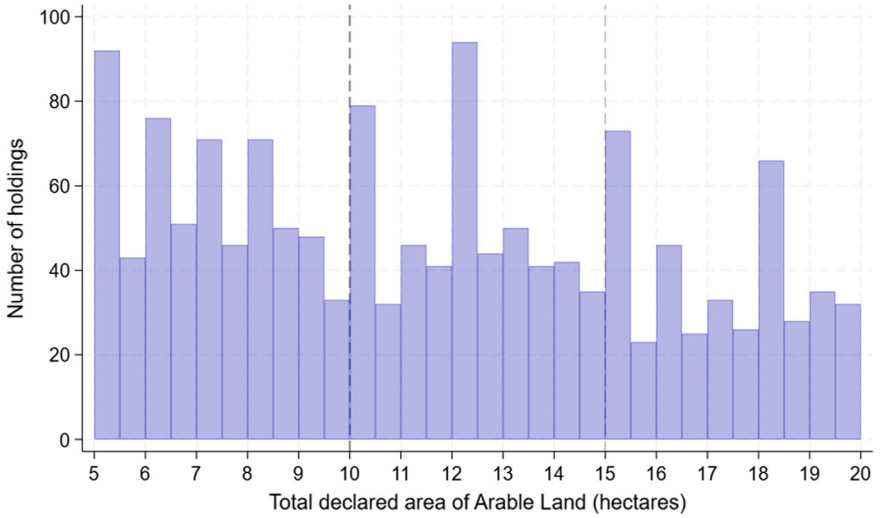


Fig. 31.4 Distribution of Spanish farm holdings by arable land size (FADN survey data, 2021). *Source* Authors' elaboration from FADN data

are clear in this context. Bringing the farm's arable land from 10.1 to 9.9 hectares or from 15.1 to 14.9 hectares exempts the farmer from the burden to comply with the relevant Greening obligations and therefore it would be reasonable to expect an out-of-pattern accumulation of holdings just under the 10- and 15-hectares thresholds, i.e. observing a behavioural response to threshold-based policies that is commonly known in Economics literature as 'bunching'.[6]

The picture changes heavily if the analyst can leverage administrative data on Spanish farms when setting up her empirical investigation. For this illustration, we have obtained data belonging to one of the high-value datasets indicated in Table 31.1 earlier in this chapter, the GSAA (Geospatial Aid Application) family. This data is routinely collected by EU Member States as part of the computerised Aid applications and Payments Claims subsystems of IACS (Integrated Administration and Control System), as defined by Art. 68 of Regulation 1306/2013. For Spain, FEGA (*Fondo Español de Garantía Agraria*) is the institution in charge of coordinating with the regional paying agencies to ensure the correct payment of CAP funding, as well as maintaining the necessary data infrastructure. Simply put, the data we use in this illustration comes from the computerised system used by farmers to declare and update, on a yearly basis, the plots of land on which they exercise agricultural activity and for which they claim CAP subsidies. The system is based on a collection of digital orthophotos that cover the entire national territory, and which farmers use as the baseline map for their digital land plot declarations. In Spain this system first came into effect in 2005, with the purpose of making it easier for farmers to submit claims, aided by graphic support, as well as making administrative and land

[6] Kleven, H. J. (2016). Bunching. *Annual Review of Economics, 8*, 435–464.

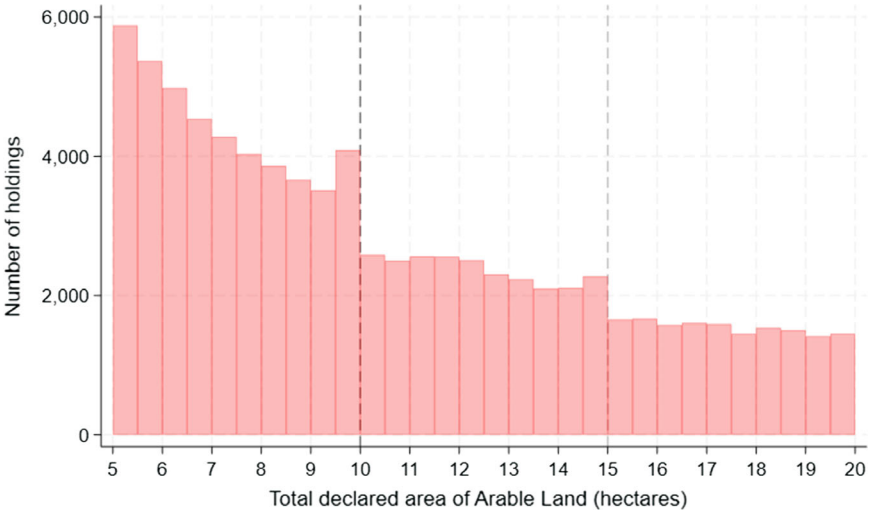


Fig. 31.5 Distribution of Spanish farm holdings by arable land size (FEGA administrative data, 2022). *Source* Authors' elaboration from FEGA data

controls easier; by design, it covers the full population of Spanish CAP subsidy claimants. In fact, data outputted by this system represents a quite ideal example of administrative data with application in environmental policy research: it is routinely and systematically collected for purposes that are unrelated to research or policy analysis and is thus available at no additional cost; it is timely and updated every year, as new CAP subsidy campaigns are rolled out; it is granular and accurate; its coverage is universal.

Figure 31.5 illustrates the result obtained when replicating with 2022 FEGA data the exercise of plotting the distribution of farm holdings declaring between 5 and 20 hectares of arable land, grouped in half a hectare size bin; the number of holdings recorded is of approximately 83,300.

It is immediately evident that the information extracted from the administrative dataset is more comprehensive and accurate than what was obtained earlier using the survey. The distribution of holdings displays an overall smooth and noiseless pattern and there is no evidence of inaccuracies in reporting, such as round-number bias; the fact that the incentives for correct-reporting that underly this administrative dataset are high and directly connected to subsidy receipt clearly shows in the final data product. Further, the fact that the universe of subsidy claimants is covered, rather than only a sample thereof, immediately ensures a correct representation of the relative frequency of holdings in the different size classes.[7] Importantly, the researcher is

[7] This result may be obtained also in survey data, if the appropriate sample weights are available. However, weights are not always available for the dimensions of interest: the FADN survey is designed to be representative in terms of region, economic size, and type of farm—not physical extension of land owned.

also able to spot the presence of the expected behavioural response to the Greening thresholds, as there are excess masses of holdings just below the 10- and 15-hectares thresholds. Strategic sorting around a policy threshold is a phenomenon that calls for appropriate research designs and corrections when estimating policy impacts, and ignoring its occurrence produces wrong evaluation results.

In conclusion, this illustration demonstrated the kind of opportunities researchers may lose when endeavouring in an environmental policy research project without having access to the relevant administrative data and having to rely on surveys instead—noting again that the survey used for this illustration is among the very best available on the EU farming sector.

4 Integrating Environmental Data from Multiple Sources

The preceding paragraphs underscore the criticality of leveraging granular data for evaluation purposes. In this paragraph we discuss the issue of integrating environmental data from different sources.

Georeferenced spatial datasets, exemplified by those illustrated in Table 31.1, represent repositories of location-specific observations, collecting 'discrete' observations from a statistical perspective, where each data point quantifies a spatial-specific aspect of an environmental outcome of interest. The LUCAS initiative maps soil attributes, pinpointing a diverse array of chemical compositions and physical characteristics, and thus providing a comprehensive picture of soil health across various geo-spatial positions. EIONET focuses on water environments, deploying a network of monitoring stations that record water quality metrics, each positioned to capture data across a spectrum of geo-spatial points. Certain gas emissions are also directly monitored using instruments such as air quality monitoring stations, industrial emissions monitoring systems, and satellite observations. This direct monitoring provides real-time or near-real-time data on emissions from specific sources. Further insights into biodiversity can be derived from sources such as the Pan-European Common Bird Monitoring Scheme (PECBMS), which presents a set of updated European common bird indicators over time and provides detailed information on several species. Further discrete data collections are the GSAA (Geographical Information System of Agricultural Subsidies) systems, which encompass farmers' applications for Common Agricultural Policy (CAP) subsidies in terms of farmland usages at the parcel level. This data offers a granular look at the intersection between agricultural management and policy support.

When conducting analyses that involve combining internal data with variables from various sources, a key challenge is deciding on the most appropriate spatial resolution for the study, meaning the selection of a spatial grid to represent the collected data. For example, in evaluating the effects of the Common Agricultural Policy (CAP) on environmental indicators like soil or water quality, subsidy information may be tracked at the individual farm level using the Farm Accountancy Data Network (FADN) or at the parcel level through administrative records maintained by

local paying agencies. However, the environmental data concerning soil and water quality are often gathered at different spatial scales and may not directly correspond to the geographic units used for subsidy data. This discrepancy creates a need to reconcile or resample the data across different spatial scales to facilitate meaningful analysis and interpretation of the effects of agricultural subsidies on environmental quality.

To address this task for analyses purposes, two fundamental solutions exist. The first involves aggregating the information into a common spatial grid with lower granularity, which is typically administrative in nature but can also be defined by other criteria, such as at the catchment level, or other. This approach simplifies the data by summarising it into a lower number of more manageable units. The second solution maintains and also increases the original high granularity of the data, calibrating the outcomes of interest to create 'pseudo-continuous' maps as a grid at a high level of resolution, composed of cells. Each cell represents a specific geographic area (often a square cell defined by its size and location) and contains the value representing a particular attribute of that area.

Both solutions have their merits and are widely utilised, depending on the specific needs of the analysis. Below, we explore some of the strengths and weaknesses of these two approaches, particularly in the context of evaluating environmental policies.

A key advantage of the first method is that the aggregated values tend to be more robust. Robustness is obtained combining multiple data points into a single, larger unit, which can reduce the impact of outliers and random variability. However, a significant drawback is that this aggregation can lead to a loss of precision. Indeed, in the context of policy evaluation, the statement 'the higher the granularity, the better the precision' is true. By reducing the granularity, the fine details of the data are averaged away, which can potentially compromise the accuracy of the analysis to the point of making it ineffective.

The second method offers the advantage of allowing analysis to be conducted with a high level of detail and providing the researcher with numerous observations. This detailed approach is crucial for analyses such as monitoring or assessing that aim to provide descriptive scenarios at high spatial resolutions. However, spatial calibration is a complex process that estimates the data points by relying on theoretical models and assumptions, as well as policy data as input. Consequently, the calibrated values derived from this approach are model-based, and therefore model-dependent. Using such data points to evaluate an environmental policy may not be optimal, since the map-based environmental outcomes may already reflect the policy's potential effects. When evaluating the Common Agricultural Policy, for instance, it is essential to ensure that the environmental outcomes maps used to measure the policy's effect are not calibrated with policy intervention data.

In conclusion, high-resolution calibration methods seem very promising for environmental monitoring and assessment exercises. However, for evaluation purposes, it is crucial to ensure that data processing avoids any biases that could potentially influence the estimation of policy outcomes. This entails attention to detail to maintain the integrity and objectivity of the evaluation process.

5 Concluding Remarks

The advancements in EU environmental data availability and the establishment of a robust legal framework for data sharing have significantly contributed to the progress of environmental research and policy analysis. The introduction of High Value Datasets and the INSPIRE data sharing portal, among other initiatives, have facilitated the accessibility and interoperability of environmental data across the European Union. However, challenges such as data harmonisation, integration, and timely availability persist, requiring continuous efforts from Member States and data-holding institutions.

The value of high-quality, administrative environmental data cannot be overstated, as it offers superior insights into environmental processes and policy impacts. High-quality data is crucial for identifying hidden environmental relationships and causal links, as lower-quality sources may not allow the detection of these connections. The acquisition of a comprehensive and accurate understanding of the situation under analysis enables policymakers to make informed decisions and develop effective strategies for environmental management.

Furthermore, the integration of environmental data from multiple sources, while maintaining the integrity and objectivity of the data, is crucial for meaningful analysis and evaluation.

As the European Green Deal continues to shape the EU's environmental policies and objectives, the importance of high-quality, accessible, and interoperable environmental data will only grow. It is imperative that the positive trend in data availability is not only maintained but accelerated, ensuring that the wealth of data generated by public and private institutions alike is harnessed to support evidence-based policymaking, contribute to the fulfilment of the Green Deal objectives and ensure the continued progress in achieving a sustainable and environmentally-friendly future for the European Union.

References

Andersson, J. J. (2019). Carbon taxes and CO2 emissions. *American Economic Journal: Economic Policy, 1*(4), 1–30.

Barreiro Hurle, J., Bogonos, M., Himics, M., Hristov, J., Perez Dominguez, I., Sahoo, A., Salputra, G., Weiss, F., Baldoni, E., & Elleby, C. (2021). *Modelling environmental and climate ambition in the agricultural sector with the CAPRI model.* Publications Office of the European Union.

Costa, H., de Rigo, D., Libertà, G., Houston Durrant, T., & San-Miguel-Ayanz, J. (2020). *European wildfire danger and vulnerability in a changing climate: towards integrating risk dimensions.* Publications Office of the European Union.

De Brogniez, D., Ballabio, C., Stevens, A., Jones, R., Montanarella, L., & van Wesemael, B. (2021). A map of the topsoil organic carbon content of Europe generated by a generalized additive model. *European Journal of Soil Science, 66*, 121–134. https://doi.org/10.1111/ejss.12193

de Rigo, D., Libertà, G., Houston Durrant, T., Artés Vivancos, T., & San-Miguel-Ayanz, J. (2017). *Forest fire danger extremes in Europe under climate change: Variability and uncertainty.* Publications Office of the European Union.

European Commission. (2020). *Sustainable Europe Investment Plan - European Green Deal Investment Plan. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions.* Brussels.

European Commission. (2022a). Commission Implementing Regulation 2023/138 of 21 December 2022 laying down a list of specific high-value datasets and the arrangements for their publication and re-use. *Official Journal of the European Union - L 19/43.*

European Commission. (2022b). SWD 195 final - Evaluation of Directive 2007/2/EC establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Commission Staff Working Document*.

European Environment Agency. (2023, September 11). *Environmental protection expenditure.* Retrieved June 14, 2024, from EEA indicators: https://www.eea.europa.eu/en/analysis/indicators/environmental-protection-expenditure-8th-eap

Finck, M., & Mueller, M.-S. (2023). Access to data for environmental purposes: Setting the scene and evaluating recent changes in EU data law. *Journal of Environmental Law, 35*(1), 109–131. https://doi.org/10.1093/jel/eqad00

IPCC. (2022). *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (P. R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, … J. Malley, Eds.) Cambridge University Press. https://doi.org/10.1017/9781009157926

Vigiak, O., Udias Moinelo, A., Grizzetti, B., Zanni, M., Aloe, A., Weiss, F., Hristov, J., Bisselink, B., de Roo, A., & Pistocchi, A. (2023). Recent regional changes in nutrient fluxes of European surface waters. *Science of The Total Environment, 858*(3), 160063.

Zelda Brutti is head of section on policy impact assessment at the Department of Health, Abu Dhabi. Before joining the DOH, she served as an economist at the European Commission's Joint Research Centre (JRC). Her expertise lies in policy impact evaluation across various sectors, including agricultural, environmental, and education topics. She spent her doctoral training between the European University Institute and at the London School of Economics, earning her Ph.D. from the EUI in 2016. Before joining the Commission, she was a researcher and lecturer in Economics at the universities of Barcelona, Pompeu Fabra, Bolzano-Bozen, and ESADE.

Marzia Freo is a professor of Economic Statistics at the University of Bologna. She served as a quantitative policy analyst at the European Commission Joint Research Centre (JRC) in the Competence Centre on Microeconomic Evaluation (CC-ME), Ispra (VA). Marzia holds a Ph.D. in Statistics from the University of Bologna and has a 20-year research career in applied statistics, micro-econometrics, and policy impact evaluation with several peer-reviewed publications.

Laura Serlenga is a professor of Econometrics at the University of Bari 'Aldo Moro' and previously served as a policy analyst at the Joint Research Centre (JRC) in the Competence Centre on Microeconomic Evaluation (CC-ME), Ispra (VA). She is also a research fellow at IZA, Bonn. In 2006, she was a Marie Curie Fellow at CREST, Paris. Laura holds a Ph.D. in Economics from the University of Edinburgh (2004) and a MSc in Economics at the University of Glasgow (1999). She has a 20-year research career in applied econometrics with several peer-reviewed publications.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 32
Administrative Data Access on Firms and Markets



Stefano Verzillo, Tomaso Duso, Francesco Quatraro, Pietro Santoleri, and Alexander Schiersch

Abstract This chapter explores how administrative microdata can significantly enhance our understanding of the links between competition, innovation and productivity. Drawing on recent empirical studies, it illustrates the analytical power of such data across several dimensions. One example comes from the use of administrative microdata to demonstrate the productivity gains associated with the adoption of digital technologies by firms. A second example illustrates how public microdata allows for more precise measurement of firm-level mark-ups, drawing on evidence from the German Monopolies Commission and demonstrating the advantages of administrative data over alternative sources. In addition, methodological challenges in innovation measurement are discussed, showing how linking patent data with administrative records can provide deeper insights into the drivers and impacts of innovation. Finally, evidence from studies utilizing privileged access to European Commission registry data is presented, revealing the effectiveness of EU R&D grant policies on firm performance. These examples illustrate also some barriers to data access and outlines key priorities for improving the availability and utility of administrative microdata for policy and research purposes.

The views expressed in the chapter are those of the authors and do not necessarily reflect the opinion od the institutions of affiliation.

S. Verzillo (✉)
European Commission Joint Research Centre, Ispra, Italy
e-mail: stefano.verzillo@ec.europa.eu

T. Duso · A. Schiersch
German Institute for Economic Research (DIW), Berlin, Germany
e-mail: tduso@diw.de

A. Schiersch
e-mail: aschiersch@diw.de

F. Quatraro
University of Torino, Turin, Italy
e-mail: francesco.quatraro@unito.it

P. Santoleri
European Commission Joint Research Centre, Seville, Spain
e-mail: pietro.santoleri@ec.europa.eu

© The European Union and The Author(s) 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_32

1 Introduction

The debate on competitiveness in Europe gained significant momentum during the 2020s, particularly in the wake of the COVID-19 pandemic and the energy crisis precipitated by the war in Ukraine. These crises not only deepened the continent's economic challenges but also highlighted Europe's lag behind the US and China, especially in critical sectors. In response, two key policy reports—overseen by Enrico Letta and Mario Draghi (Draghi, 2024; Letta, 2024)—have been commissioned to propose solutions to Europe's 'competitiveness problem' to be implemented by the new European Commission that took office at the end of 2024. These reports, alongside a range of other studies, indicate that one of the primary drivers of Europe's declining competitiveness is weak productivity growth relative to the US and other global regions (see Draghi, 2024, p. 12). This productivity gap is largely attributed to the underdevelopment of total factor productivity (TFP).

In this debate, the productivity gap between the EU and the US is largely attributed to the insufficient development of the technology sector and the slow adoption of digital technology. This gap is expected to widen with the development of artificial intelligence (AI), a field currently dominated by US and Chinese companies. It is also attributed to insufficient investment in innovation. European companies appear to invest significantly less in research and development, spending half as much as their US counterparts as a percentage of GDP (Draghi, 2024).

In addition, the EU faces challenges in effectively translating basic research into commercial applications (Czarnitzki et al., 2025). Another critical issue is enabling companies to scale up, which requires facilitating smoother operations across the single market. For example, digital infrastructures such as broadband, mobile communications and cloud computing would benefit from a single European regulatory framework. Finally, while the two reports highlight the importance of fostering a competitive market environment and of completing the single market, they also suggest that Europe needs to adapt its competition policy framework to ensure that the continent remains competitive on the global stage.

A robust fact base analysis grounded on high-quality data is essential to assess the existence of a European 'competitiveness problem', analysing its drivers, and formulating effective policy responses. Understanding the forces behind TFP growth beyond the aggregate level necessitates the use of rich and timely administrative firm-level data. This is essential for a precise productivity estimation and, especially, for uncovering causal links between policy measures and interventions. For instance, high-quality, disaggregated data allowed researchers to demonstrate that enhanced research and development (R&D) efforts significantly boost TFP (e.g. Doraszelski & Jaumandreu, 2013), that investment in various intangible assets is a key driver of productivity (e.g. Le Mouel & Schiersch, 2024) and that the adoption and use of new digital technologies are important for productivity improvements e.g., (Duso et al., 2025).

Moreover, the use of firm-level, instead of aggregated, data has revealed the substantial heterogeneity in productivity among firms, even within narrowly defined

sectors. For example, not all firms benefit equally from R&D investment or innovation, which contributes to a growing gap between a small number of highly successful firms and the larger majority that see only marginal productivity gains from such investments (e.g. De Ridder, 2024; Le Mouel & Schiersch, 2024).

Similarly, the call for changes to the competition policy framework in Europe, particularly in the Draghi report, is based on the premise that such changes will allow for a more efficient allocation of scarce resources, which can then be effectively used to boost innovation and technological progress. In order to gain insights into the competitiveness of the European economy and the overall allocation of resources, it is therefore important to better understand the role of competition and, perhaps even more challenging, come up with good measurement of it. Since the mid-2010s, both academic literature and policymakers have recognized that increasing concentration and rising mark-ups in many sectors of the economy may pose a competition problem, as they can affect the economic welfare of both consumers and producers.

The relationship between competition, innovation and productivity in Europe remains insufficiently understood, in part due to limitations in the measurement of innovation. There is a pressing need to enhance the quality and granularity of innovation indicators, particularly through the more effective use of administrative data. While the empirical literature focusing on the US has rapidly grown in recent years, evidence across Europe is more scant, given the absence of a unified system to collect administrative data. It is important to deal with this shortcoming, as large-scale micro-econometric analyses may provide unprecedented contributions to the investigation of key theoretical expectations. For example, Italian data show that innovation seems not to be a distinctive feature of market leaders, while political connections matter more (Akcigit et al., 2023). US analyses on innovative start-ups instead reverse the common tenet that young start-uppers are the most successful, showing that success accrues mostly in mature entrepreneurs (Azoulay et al., 2020). Empirical results obtained with this kind of data are crucial to guide policymakers and stakeholders decisions.

This chapter gives examples of insights gained by use of administrative microdata. Specifically, Section 2 presents the study by Duso et al. (2025), which demonstrates how administrative microdata helps to understand the importance of digital technologies for higher productivity of firms. Section 3 shows how public microdata helps to measure mark-ups and presents the findings of the German Monopoly Commission using such data (Monopolkommission, 2022, 2024). Moreover, it shows how the use of high-quality administrative data can significantly improve the understanding of mark-up dynamics compared to the use of other commercial data sources. Section 4 discusses the methodological underpinnings of innovation measurement and shows how the matching of traditional indicators like patents with administrative data can shed new light on the economic drivers and impacts of research and innovation dynamics. Section 5 presents recent examples of how privileged access to European Commission registry data has allowed one to evaluate the impact of selected EU initiatives—namely R&D grants—on firm-level outcomes, thereby enhancing understanding on effectiveness of EU policies. The final section discusses existing

problems of data access and data usage and formulates possible objectives for the future.

2 Digitization and Firms' Productivity

The Draghi report stresses the importance of the digital transformation and how Europe is lagging behind in this domain. Before presenting an example of the benefits of administrative firm-level data, the estimation of total factor productivity (TFP) is briefly outlined.

The main advantage of TFP is that, unlike labour productivity, it takes into account many input factors and is therefore not as easily misleading as labour productivity. However, TFP as such cannot be observed but needs to be estimated. One of the main problems in estimating it is the well-known simultaneity problem, first highlighted by Marschak and Andrews (1944), which complicates the estimation of production functions.[1]

The severity of this problem, particularly when it comes to evaluating policy measures and their impact on productivity, can be seen by looking at a simple estimating equation for a production function as described in equation (32.1). Accordingly, log-value added (y_{it}) of firm i at time t is generated by a combination of log-labour (l_{it}) and log-capital (k_{it}) by means of the production function $f(\cdot)$. In empirical analyses, other regressors ($\mathbf{X}$) are often included to capture further differences between firms, such as sectoral-, time- or region-invariant characteristics. The simultaneity issue arises because TFP (ω_{it}) is never observed in real-life firm-level data and ends up being part of the error term (ε_{it}) when estimating equation (32.1). This in turn causes the error term to be correlated with the regressors, especially labour and capital inputs, making them endogenous.

$$y_{it} = f(l_{it}, k_{it}; \boldsymbol{\beta}) + \boldsymbol{X'_{it}\theta} + \underbrace{\varepsilon_{it}}_{(\omega_{it}, \epsilon_{it})} \tag{32.1}$$

This means that TFP, which is often calculated as a residual, is biased. Consequently, any analysis of the drivers of TFP, be it firm decisions, policy interventions or the like, may lead to inaccurate conclusions if the endogeneity problem is not addressed. Control function approaches not only attempt to overcome the simultaneity issue, but also make it possible to relate the variables of interest directly to TFP without having to estimate TFP first. They also allow the estimation of mark-ups at the firm level without the need to observe prices or marginal costs at the firm or product level.

A key concept of these approaches is the assumption that an investment or material demand function with certain characteristics exists, such as the assumption of strict

[1] For an overview of the different approaches taken in the past, see Ackerberg et al. (2007).

monotonicity between material or investment and productivity.[2] These characteristics allow the function to be inverted so that the unobserved TFP becomes the dependent variable ($\omega_{it} = h_t^{-1}$). Of course, TFP is still unobserved. However, the inverted function (h_t^{-1}) only contains observable variables and can therefore be substituted into the production function, thereby controlling for unobserved TFP, which is then no longer in the error term and can no longer cause endogeneity of labour and capital. Because the true functional form of (h_t^{-1}) is unknown, it is treated nonparametrically (Ackerbergetal, 2015). Furthermore, the variable that is assumed to drive TFP—broadband in the case of the subsequent analysis (b_{it})—must also be included already at this stage (De Loecker & Warzynski, 2012). Equation (32.2) illustrates this, with

$$y_{it} = \underbrace{f(l_{it}, k_{it}; \boldsymbol{\beta}) + h_{it}^{-1}(m_{it}, l_{it}, k_{it}, b_{it})}_{\Phi_t} + \boldsymbol{X_{it}'\theta} + \epsilon_{it} \tag{32.2}$$

The estimation of equation (32.2) is a necessary step but, unfortunately, does not provide consistent coefficients for labour, capital and the like.[3] In order to estimate consistent coefficients, control function approaches make the assumption that the development of TFP follows a first-order Markov process. As Doraszelski and Jaumandreu (2013) and others have shown, this function can be modified to include an exogenous factor that has a direct effect on productivity; b_{it} in the current example.

$$\omega_{it} = E[\omega_{it}|I_{it-1}, b_{it}] + \xi_{it} = g(\omega_{it-1}, b_{it}; \boldsymbol{\gamma}) + \xi_{it} \tag{32.3}$$

The final formulation is derived by focusing on the core and rearranged production function (32.2) and substituting it into equation (32.3), while replacing $\Phi_t(\cdot)$ with the observed $\hat{\Phi}_t(\cdot)$ from the first stage. Note that by estimating equation (32.4), both the production function coefficients ($\boldsymbol{\beta}$) and the coefficients of the law of motion ($\boldsymbol{\gamma}$) are estimated simultaneously. The resulting coefficients for labour and capital are unbiased and one can directly estimate the effect of a certain policy b_{it} (like a subsidy, a change in regulation, etc.) on TFP:

$$\hat{\Phi}_t(l_{it}, ...) = f(l_{it}, k_{it}; \boldsymbol{\beta}) + g\left(\hat{\Phi}_{t-1} - f(l_{it-1}, k_{it-1}; \boldsymbol{\beta}), b_{it-1}; \boldsymbol{\gamma}\right) + \xi_{it} \tag{32.4}$$

Such a method has been used in the analysis of Duso et al. (2025) on the effect of broadband availability on TFP. The starting point for that analysis is the general understanding that modern broadband networks are essential for many new digital solutions and business models and are just as indispensable to national economies as a modern transport infrastructure. The introduction of new network technologies is often supported through subsidies provided by national authorities. The repeated

[2] For more details see Olley and Pakes (1996), Levinsohn and Petrin (2003), Ackerberg et al. (2007), Van Beveren (2012), Ackerberg et al. (2007) and Ackerberg et al. (2015). In case of a simple production function, the material demand function is $m_{it} = h(\omega_{it}, l_{it}, k_{it})$.

[3] For reasons of space, we refrain from a detailed discussion of the reasons for this and refer interested readers to Ackerberg et al. (2007, 2015) and Van Beveren (2012).

use of public money for the roll-out of new generations of broadband technologies and the actual benefits of this are increasingly being questioned in the public debate.

To feed the public debate with facts, the study examines a past technological leap, the roll-out of broadband with a transmission rate of up to 16 Mbit/s in Germany, with the aim of identifying its causal effect on the development of total factor productivity, highlighting possible differences between sectors and drawing conclusions for the ongoing roll-out of the next generation broadband network.

The study is able to make use of a very large dataset for its analysis. It consists almost entirely of different firm-level dataset from the German Statistical Office, covers the period 2010–2015 and contains around 450,000 observations from almost all sectors.

Download speeds of 16 Mbit/s in the period under review (2010–2015) were mainly achieved by ADSL2+. Although this technology was already developed in the mid-2000s, its actual availability in Germany at the beginning of the observation period was quite modest. In 2010, it averaged between around 11% in Schleswig-Holstein and 32% in North Rhine-Westphalia across all municipalities. A detailed descriptive analysis also shows that the availability of ADSL2+ was not only very unevenly distributed across federal states but also across municipalities, with many regions where it was not available at all at the beginning of the observation period. For example, half of the municipalities in Rhineland-Palatinate had no networks with ADSL2+ technology in 2010. In a further quarter of the municipalities (median to third quartile) the share in all connections was below 3%. This means that in 2010, the networks of around three quarters of all municipalities in Rhineland-Palatinate were not yet equipped with ADSL2+ technology.

Until the end of the observation period, availability increased significantly, reaching an average of 68% in Hesse and 32% in Schleswig-Holstein, to name the upper and lower ends of the distribution. The heterogeneity between municipalities remains very high. For example, the availability of ADSL2+ in 2015 was only 4% or less in a quarter of the municipalities in Rhineland-Palatinate (first quartile). At the same time, there were a similar number of municipalities in that very same state (third quartile) where 93% or more of all lines were using ADSL2+ technology. A similarly pronounced heterogeneity can be found in the other states.

In summary, the take-up of ADSL2+ was very low at the beginning of the observation period and increased steadily over time, but without ADSL2+ becoming the dominant technology in all municipalities compared to older alternatives. This is an advantage for the empirical analysis as it provides sufficient variation to identify possible productivity effects of the introduction of ADSL2+.

Applying the above-described econometric technique the study examines how the roll-out affected the TFP of firms in every sector of the German business economy. Figure 32.1 contains the coefficients of broadband availability, which are depicted as points, while the whiskers represent the 95% confidence intervals. Note that both the results for one-digit industries and for two-digit industries are included in the figures. The former are depicted using red dots and whiskers while the latter are shown in black. In case the coefficients are not significantly different from zero, they are depicted in grey and are not filled.

When focusing on the one-digit industries, one sees that coefficients for broadband are significant in all one-digit sectors and the confidence intervals are very small. As the study goes on to show, these coefficients are smaller than those estimated by regressing broadband connectivity on output. This indicates that the productivity-enhancing effects of a new broadband technology are smaller than the overall benefits of the new technology. In other words, higher productivity accounts for only a fraction of the overall positive effects that are generally found for a new broadband technology.

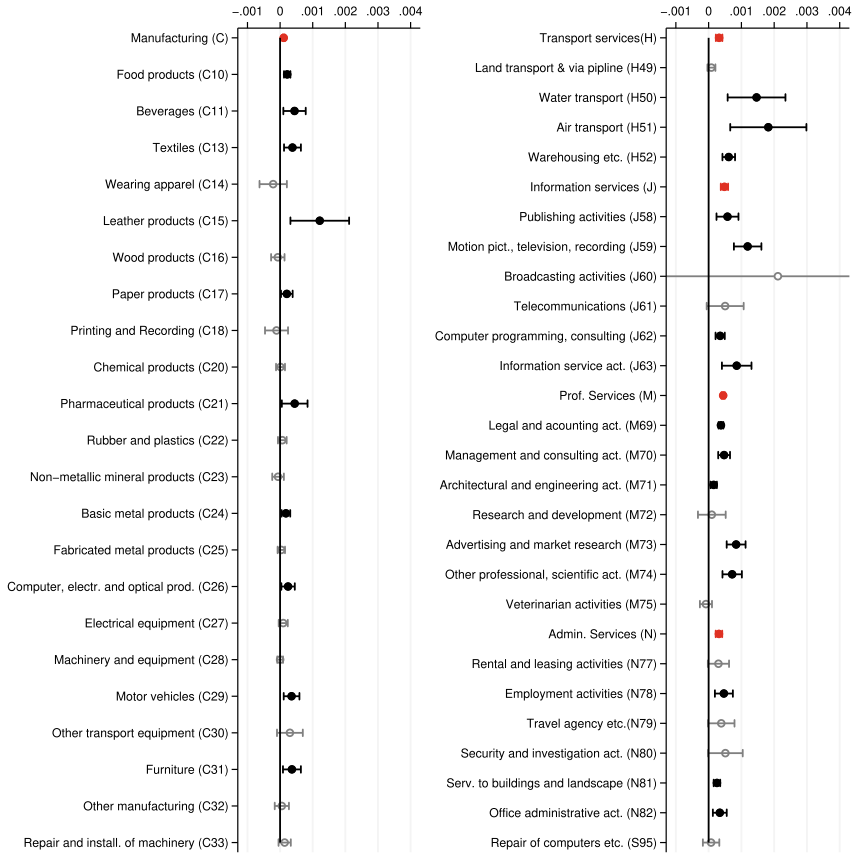
Importantly, the results already indicate that the effects vary at the level of broadly defined sectors. Many studies have to stop at this level of detail, if they are able to look separately at the 1-digit industries at all, because the data are often not large enough or good enough to perform a detailed analysis by industry. In contrast, Duso et al. (2025) were able to analyse the effects below the level of one-digit industries thanks to excellent public microdata. This makes it possible to uncover differences between very specific industries. For example, do firms in the automotive industry benefit from a new broadband technology to the same extent as firms in the machinery industry?

Figure 32.1 demonstrates that the increased availability of 16 Mbit/s had no significant impact on the total factor productivity of firms in more than half of the industrial sectors. This is a very different result from the one for one-digit industries, as one can no longer simply argue that broadband generally increases the total factor productivity of manufacturing firms. By contrast, the majority of firms in the various services sectors can benefit from the improved availability of ADSL2+ in terms of increased TFP. This applies equally to firms in the different sectors of transport services (ISIC code H), information and communication services (ISIC code J) and business services (ISIC codes M and N). Considering only the significant coefficients, the average increase in ADSL2+ availability of around 38 percentage points over the observation period led to an increase in productivity of between 0.8% and 6.8% in the respective sectors. Taking into account the corresponding confidence intervals, the maximum increase in productivity was up to 13.1%. This order of magnitude is within the range of effects of R&D investment reported in the literature.

However, the large heterogeneity of the coefficients also demonstrates that firms in different sectors benefit very differently from the introduction and expansion of the new broadband technology. Moreover, service firms seem to benefit the most, while the effects are much smaller for manufacturing firms.

The study also examines whether a dense fibre optic network is preferable. To this end, the technical features of ADSL2+ are exploited. The ADSL2+ technology and the period under consideration are characterized by the fact that the ‘last mile’ to the customer is primarily made up of copper cables. Due to the properties of copper, the transmission volume and the transmission rate decrease as the length of the copper line increases.

The study capitalizes on this fact and investigates whether productivity-enhancing effects can be found regardless of the distance to be bridged with copper cables. It finds that the increased availability of ADSL2+ has a primarily positive effect on TFP if there is a dense fibre optic network with short distances to the end customer. In the case of a thin network and long distances bridged with copper lines, there are



(a) Manufacturing (b) Service sectors

Fig. 32.1 Coefficients and confidence intervals for the effect of broadband on TFP. The wiskers are calculated such that they match the 0.05 significance level. Based on findings from Duso et al. (2025)

hardly any statistically significant effects. This means that a sufficiently dense fibre optic network must be available in all areas, both rural and urban, in order for the productivity-enhancing effects of broadband networks to occur at all.

Although the analysis focuses on the roll-out of ADSL2+, general conclusions can be drawn from the results with regard to the further roll-out of new broadband technologies. First, from an economic point of view, continued support for new roll-outs is generally the right thing to do. Many firms and industries benefit in the form of higher TFP and can thus increase their competitiveness. This in turn is a prerequisite for maintaining or even increasing employment.

However, the results show considerable variation across sectors. Policymakers should therefore not assume that the use of new technologies will benefit all sectors

equally. At least in the case of ADSL2+, it is mainly the service sectors, especially information and communication services, where network deployment has enabled productivity gains.

Secondly, productivity gains can only be expected if the network is designed in such a way that, if direct fibre optic connections are not possible, the distances to be bridged with copper cables must be as short as possible. For example, the supervectoring technology, which can achieve transmission rates of 200 Mbit/s, requires that the distances to be bridged with copper cables are no longer than 400 metres. Therefore, politicians would be right to support the rapid conversion of existing networks and the build-up of new fibre optic networks in unconnected regions. This is clearly necessary and can be defended in the public debate on the basis of these findings, which show that without it, businesses in these regions will not benefit from new broadband technologies and will fall further behind in their competitiveness, exacerbating the structural weakness of rural areas.

3 Mark-ups

A central issue in current policy discussions is the relationship between competition and competitiveness. Specifically, does a more competitive environment, supported by robustly enforced competition policy, enhance or hinder competitiveness? Answering this question necessitates accurately measuring competition. A common metric used to assess the level of competition is the mark-up, denoted as μ_{git}, which is typically calculated as the ratio of the market price p_{git} of good g, produced by firm i at time t and its marginal cost MC_{git}. The prevailing assumption is that in a perfectly competitive market, firms set their prices equal to marginal cost. However, firms with market power can sustain prices above marginal cost in the long run. Thus, a positive mark-up can be seen as a measure of the market power of that particular firm i in the market for product j.[4]

While this measure of competition is theoretically sound, possibly even more so than other measures traditionally used in research and policy analysis, such as concentration indices, its empirical measurement is challenging. One of the problems is the limited availability of relevant data needed to estimate mark-ups. In particular, while data on prices are rarely available, data on marginal costs are almost never accessible. This can be overcome by estimating complex structural models of supply and demand. However, the estimation of mark-ups on the basis of a full model of competition can often be cumbersome, especially if one wants to go beyond the estimation of mark-ups for a single product. This is relevant because, for policy

[4] The sources of market power can be diverse and have different implications. Product differentiation and multi-product ownership are important aspects. Firms' behaviour, i.e. the extent to which they react strategically to each other, is another e.g., (Nevo, 2001); (Slade, 2004).

purposes, one is often more interested in average mark-ups at the firm level than in very specific and precise mark-ups at the level of a single product.[5]

To address these issues, a recent literature has emerged that derives a different mark-up relationship from cost minimization (DeLoecker & Warzynski, 2012). Accordingly, a firm's mark-up can be expressed as the ratio of the output elasticity β^x of any flexible input x to the input's cost share in total sales α^x. Thus:

$$\mu_{it}^{x} = \frac{\beta_{it}^{x}}{\alpha_{it}^{x}} \tag{32.5}$$

This production-side approach allows the estimation of firm-level mark-ups on the basis of balance sheet data from a few variables that are much more readily available. While cost shares can be derived directly from balance sheets, output elasticities are generally not observed directly. However, as we saw in the previous section, they can be estimated using econometric techniques. Indeed, any production function, such as the one shown in Eq. (32.1), allows, under certain assumptions, to identify $\beta_{it}^{x} = \frac{\partial f(\cdot;\beta)}{\partial x_{it}}$, where the flexible input x often used in the literature to identify mark-ups is labour.

This approach has now become quite standard in the empirical literature not only in industrial organization but also in macroeconomics and trade. For instance, based on this method, the now seminal paper by De Loecker et al. (2020) estimates mark-ups for the US economy over a long period of time and across industries and analyses the macroeconomic implications of these trends. They document a significant increase over time driven mainly by the upper tail of the mark-up distribution. Calligaris et al. (2018, 2024) show that similar, though much less pronounced, rising trends can also be observed for Europe during the last decades. These findings have sparked heated debate on the reasons for these mark-ups trends and the appropriate policy response.

The main purpose of this section is to highlight how the availability of high-quality administrative firm-level data affects the quality of the estimates. This is not only crucial for identifying and understanding the sources of the potential problem, but perhaps even more importantly, for designing the right instruments to address it.

Depending on the research question, the quality of estimates and their representativeness for a population depends on the quality and representativeness of the underlying data. The studies mentioned above estimate mark-ups for the entire economy based on data from commercial sources. De Loecker et al. (2020) uses data from Compustat, while Calligaris et al. (2018, 2024) use Moody's Orbis database. Although these datasets are of high quality and have been extensively used in research, they are not perfect. For one thing, they are not representative. The former only covers listed companies. As a result, the database is skewed towards the largest companies in the economy and accounts for only 29% of private employment in the US. The Orbis database is probably more representative of the economy as a whole, as it also covers small and micro enterprises. However, there are large differences in coverage between countries as well as in the quality of the reported data,

[5] For an interesting discussion of these issues, see for instance Berry et al. (2019).

which are often missing for some countries and variables.[6] Germany, for example, is relatively poorly covered. In addition, the fact that they are based on publicly available information rather than mandatory self-reporting introduces an additional sample selection problem that makes these datasets less representative. Finally, both databases contain little information compared to administrative data and lack richer variables that could be important for modelling a firm's production function.

The use of administrative data can be very useful in this context. An example is the work of the German Monopolies Commission (GMC)—an independent committee of experts that advises the German government and legislature on competition and regulation issues. For decades, the Commission has reported on the development of various measures of competition in Germany, including more recently mark-ups, in the traditional chapter on the state of competition in its biannual report. For several reasons that will be discussed below, here we focus on the 2022 report (Monopolkommission, 2022).[7]

Similar to the studies cited above, the GMC uses the framework of De Loecker and Warzynski (2012) to estimate mark-ups, among other measures of competition. As discussed above, this requires an estimate of the production function and, in particular, of the output elasticity of a flexible input. The Commission follows the extensive literature and the method discussed in the previous section and implements the two-stage control function approach by Ackerberg et al. (2015).[8] As explained above, the firm's material input is used to resolve the simultaneity problem inherent in the estimation of a production function, and helps identify and estimate firm-specific productivity. Therefore, labour is used as the potentially variable input factor to identify the mark-up according to equation (32.5).[9]

Unlike previous studies, the GMC primarily relies on high-quality administrative data from the AFiD[10] panels, as outlined above. To highlight the importance of using a more comprehensive and representative data source, they also incorporate Orbis data, comparing how the choice of dataset affects their estimates. This comparison is a key focus of this section, underscoring the significance of data quality in the analysis.

Both the arithmetic and the turnover-weighted mean of the estimated average mark-ups for manufacturing and services are shown in Fig. 32.2. In the manufacturing

[6] See the excellent discussion by Bajgar et al. (2020) on the quality of the Orbis database.

[7] For 2022, the chapter has been translated in English and can be downloaded here. The newest estimates are contained in the 2024 report (Monopolkommission, 2024).

[8] Similar to the previous chapter, a specification of the production function is modelled in which gross output in euro is explained by the three input factors capital, labour and a composite variable for flexible factors of production, e.g. material input. In the absence of firm-level price data, and in line with the extensive literature, two adjustments are made. Firstly, monetary variables are deflated with price indices at the industry level, which controls for price changes in the industry average. Second, firm-specific and time-varying wages and market shares are used to control for cross-firm differences.

[9] This is not an innocuous assumption, which has been discussed and assesses in the report.

[10] The AFiD acronym stands for the German 'Amtliche Firmendaten für Deutschland', the Official Firm Data for Germany.

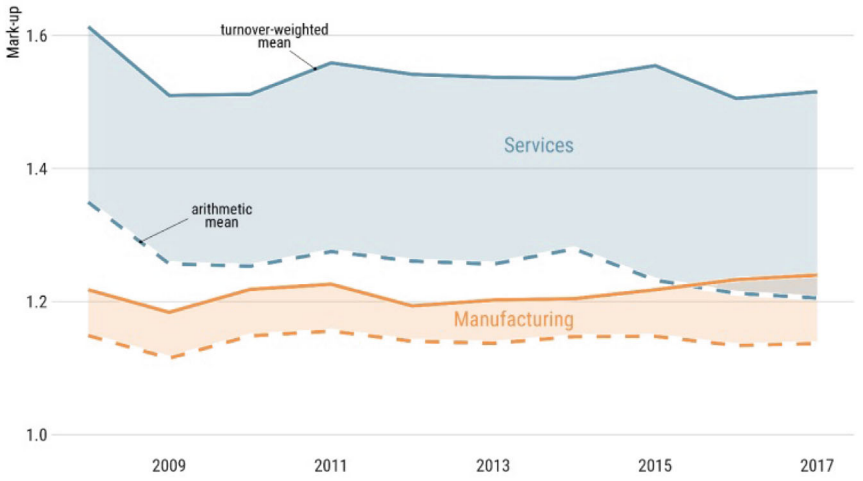


Fig. 32.2 Estimated Mark-ups for Germany. *Source* Bi-annual report of the German Monopolies Commission (Monopolkommission, 2022)

sector, mark-ups have been fairly constant, with a slight increase since 2013 and the lowest point in the 2009 financial crisis, when weighted average mark-ups were below 20%. Mark-ups in services, although higher than in manufacturing, appear to have fallen first around the time of the financial crisis and then significantly from 2015 onwards. Compared to estimates for the US and other European countries, mark-ups in Germany appear to be significantly lower. Moreover, the rising trends observed for other countries are not observed for Germany.[11]

As in other studies, the GMC notes that mark-ups appear to be positively correlated with both firm size and concentration. Indeed, large firms in services have much higher mark-ups than smaller firms, and they decrease much less over time. Moreover, mark-ups in services appear to have increased significantly for markets in the highest quartile of the concentration distribution. In addition, further analyses show that sectoral developments in mark-ups are mainly driven by firm-specific changes and increased turnover of firms with high mark-ups rather than by shifts in market shares between firms.

An interesting exercise reported by the GMC in its 2022 Biennial Report was to assess how the source and quality of the data affected the main results. Comparing mark-up estimates from administrative data with those from private databases, such as Orbis, can help to assess the validity and representativeness of such datasets.

[11] The report carefully assesses the many assumptions needed to estimate the production function and hence the mark-ups. For example, it replicates the results assuming a different functional form for the production function, assessing the role of labour adjustment costs or using different flexible inputs to derive the mark-ups. While some of these changes appear to affect the reported results, especially the observed trends, more than the levels, are quite robust.

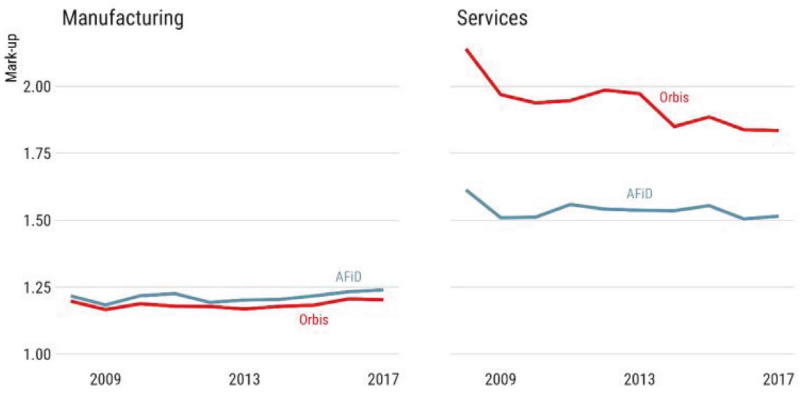


Fig. 32.3 Estimated Mark-ups for Germany—comparing administrative to commercial data. *Source* Bi-annual report of the German Monopolies Commission (Monopolkommission, 2022)

Orbis is particularly important for European studies. It is very widely used not only in academic research but also in policy work because it is probably the most complete cross-country firm-level database and since it appears to be more representative of the overall firms' distribution than other private databases. However, it is well known that even Orbis still under-represents small firms, which can bias estimates of sector-specific mark-ups, especially if firm size is correlated with mark-ups. Thus, estimates based on administrative data can serve as a benchmark to assess the potential bias that would be introduced by using lower quality data sources.

The results of the comparison of aggregate mark-ups estimated with the AFiD panel and Orbis data, shown in Fig. 32.3, are quite striking and informative. While both the level and the trends of the estimated mark-ups are very similar across the databases in manufacturing, the differences in the service sectors are substantial. The mark-ups estimated using Orbis data are much higher and show a much sharper decline over time than those estimated using administrative data, which is likely to reflect sampling bias due to lower firm coverage, higher variance and the influence of firm size. This is a very good example of the importance of basing policy-relevant analyses, such as the Monopolies Commission's assessment of the state of competition, on high-quality administrative data rather than on commercial databases.

Another advantage of the stratified and randomly sampled administrative firm microdata over other data sources is to provide high-quality, representative information on specific sub-groups of firms. This allows a closer examination of individual industries within the manufacturing and services sectors and helps to identify different trends in mark-ups (Fig. 32.4). Again, such analyses are of policy relevance as they allow potential policy responses to be targeted at those sectors that appear to be most problematic.

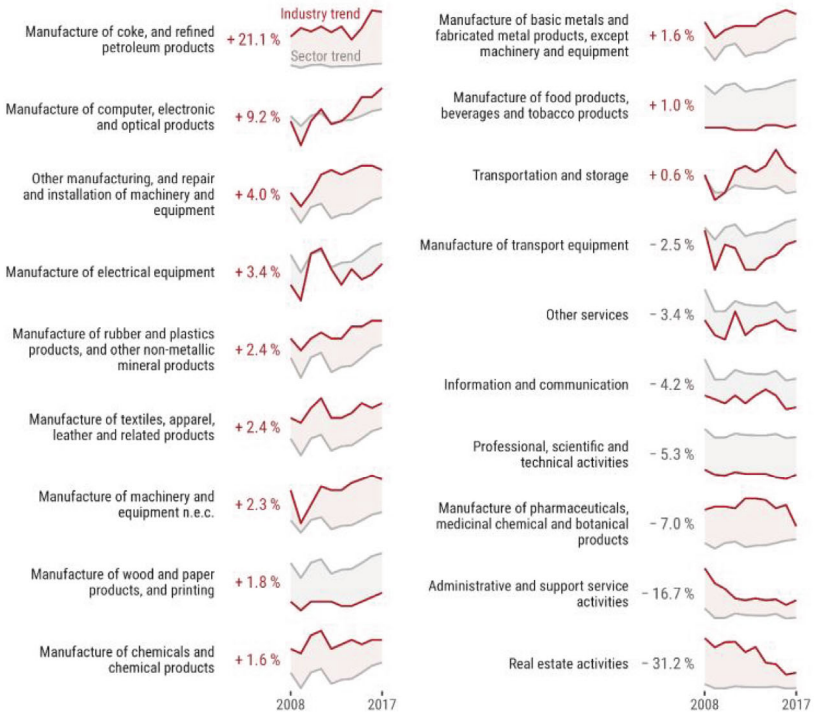


Fig. 32.4 Estimated mark-ups for Germany—comparing administrative to commercial data. *Source* Bi-annual report of the German Monopolies Commission (Monopolkommission, 2022)

The heterogeneity in the mark-ups levels across industries within sectors appears to be quite large. For example, the manufacturing of machinery and equipment shows mark-ups way above the manufacturing sector average, while professional, scientific and technical activities display below-average mark-ups. Moreover, the analysis highlights significant differences across industries also in the mark-up trends. Notably, the manufacture of coke and refined petroleum products experienced the largest increase in mark-ups, starting from a level already 30% higher than the manufacturing average in 2008, with a particularly sharp rise after 2014. In contrast, among service industries, only transportation and storage saw an average mark-up increase, while the rest followed the sectoral downward trend, with the most significant drop occurring in real estate activities.

In summary, this section shows the importance of the availability of high-quality, representative firm-level data for both research and policy purposes. We focus on the assessment of the state of competition, and in particular of mark-ups, by a policy institution, the GMC. While this illustrates the importance of access to high-quality data, it may come at a price. For one thing, the data provided by the statistical offices are not easily accessible and downloadable. It can only be accessed in the data room

of the statistical office and the data can never really be seen and tested. Every single result, whether it is a figure, a table of descriptive statistics or a regression table, has to be checked and approved by officials at the Statistical Office. But perhaps the most problematic aspect, especially for using the data for policy purposes, is the fact that the data come with a very long time lag of several years. This is mainly due to the very thorough quality checks carried out by the statistical offices, but also to limited resources. This is very problematic as policymakers often have to deal with very urgent policy issues for which up-to-date data are crucial, such as during the COVID-19 pandemics or the recent energy crisis in 2022.

A concrete example is again provided by the biennial report of the GMC 2024 (Monopolkommission, 2024). A sub-section of the chapter on the state of competition deals with a particular issue that was hotly debated in the inflationary period following the energy crises triggered by the war in Ukraine: the so-called 'seller's inflation' or 'greedflation' debate. The policy question was to what extent firms would take advantage of the general rise in price expectations to raise prices disproportionately and thus become a further driver of inflation. Following the academic literature e.g., (Conlon et al., 2023) to assess this question, the GMC proposed a descriptive comparison of the development of mark-ups and the development of price indices. While a positive relationship between rising prices and rising mark-ups could have many causes, it would at least be consistent with the seller's inflation hypothesis.

For this highly policy-relevant analysis, the GMC unfortunately had to rely on the Orbis dataset instead of the higher-quality AFiD panel for two main reasons. First, the most recent data from the statistical offices, published in 2024, only extended to 2021. Second, the administrative data do not cover the food retail sector, a key area for understanding inflation dynamics. Based on the Orbis data, supplemented by publicly available aggregated data for the agricultural sector from the Federal Ministry of Food and Agriculture, the GMC found no evidence to support the 'seller's inflation' hypothesis. Instead, the observed correlation patterns are consistent with incomplete cost pass-through. Pass-through rates are lower in manufacturing than in services, and do not appear to have changed around the 2019–2022 inflation period.

While obstacles exist in the real world, we also believe that they are not related to the administrative nature of data, but rather to data access procedures, as further discussed in Chap. 33.

4 Innovation

The emphasis on the importance of digital technologies for productivity growth and firms' competitiveness calls for a broader discussion on the crucial role of innovation in these dynamics. In Economics, Solow's contribution has paved the way for a fertile debate on the impact of technological change and innovation on TFP (Solow, 1957). As stressed by Griliches (1994), the discovery of the residual has initially led economists to interpret it as a measure of technical change. Yet, it was clear that such an interpretation was partly arbitrary, and more efforts were needed to understand the

origins of the residual. The former approach to solving this problem revolved around the hypothesis that the observed differences between real output growth and real input growth were due to more or less severe measurement errors. Accordingly, research efforts have been dedicated to refining the measurement of capital services and, in general, to reduce any possible source of biases in accounting procedures, with the expectation that any improvement in this direction should have led to the reduction of the observed residual (Griliches & Jorgenson, 1966; Jorgenson & Griliches, 1967).

Yet, while advancements in the theory of capital and its measurement have significantly refined the calculations of total factor productivity, they have not proved to really explain the residual away. The research programme started by Zvi Griliches at Harvard University in the 1970s was actually aimed at finding explanations of the residual by attributing it to specific sources. Besides the improvements in the quality of labour and capital services, empirical studies have found that efficiency gains and technical change are responsible for productivity growth and that R&D and innovation are the main responsible for these dynamics (Griliches, 1994).

The former empirical setting for the estimations of the productivity impact of R&D expenditure has been put forth by Griliches (1979), who proposed the adoption of the so-called *extended production function.* The generic form for such an equation is as follows:

$$Y = F(X, K, u) \tag{32.6}$$

where Y is a measure of firms' output, X is an index of conventional inputs, K is a measure of the current state of technological knowledge and u captures all other unmeasured determinants of output. The current state of knowledge is, in turn, a function of present and past investments in research and development activities, which is obtained by applying a distributed lag function to R&D expenditures R to obtain a sort of stock measure. Let $W(B)$ be a polynomial of the lag operator B, the function to estimate technological knowledge takes the following form:

$$K = G[W(B)R, v] \tag{32.7}$$

where v is a technological shock.

Substituting Equation (32.7) in Equation (32.6), one obtains a functional form explicitly connecting R&D investments to firms' output and productivity. Equation (32.7) formalizes the concept of R&D capital stock or knowledge capital stock. Yet it leaves some measurement issues open as to (i) which is the most appropriate lag structure $W(B)$ accounting for the delayed economic impact of R&D expenditures, (ii) which is the most appropriate depreciation scheme to be applied and (iii) how to estimate the economic impact of intersectoral R&D spillovers.

There are major problems associated with the use of R&D statistics. First of all, they are an input measure, which can translate into different outputs. The use of R&D data hence implies an aggregation of these different outputs, with different economic values, in one single measure. Second, the distribution of R&D expenditure is strongly asymmetric across sectors and firms' size, suggesting that these

data are better suited to capture innovation efforts in large high-tech firms. Third, the composition of R&D expenditure is heterogeneous, and it is important, though very difficult due to the available data, to distinguish between basic and applied, or corporate vs. government-funded R&D. Fourth, and perhaps most importantly, the reporting of R&D expenditures significantly varies across countries. This raises a major issue of comparability in large international datasets, both at the micro- and at the macro-level. Moreover, the heterogeneity of economic incentives to invest in R&D across countries is likely to engender a reporting bias in countries with the most generous schemes (Griliches, 1979, 1990; Mansfield, 1981; Pavitt, 1982).

In view of these shortcomings, the search for a reliable indicator of economically valuable technological knowledge has led to the identification of patents as the most popular measure of innovation output. The link between R&D expenditures and patents is well represented by the so-called *knowledge production function* (KPF), which has been conceptualized by Pakes and Griliches (1984).

The KPF provides a framework to model the production of knowledge (K) as the result of the translation of R&D expenditures (R) into inventions (P). Empirical analyses have actually provided support to this model contributing to the popularity of patents as indicators of innovation outcome (Hall et al., 1986; Jaffe, 1986). Combining the KPF with the extended production function (Equation 32.6), one obtains a system of equations which is at the basis of the model proposed by Crépon et al. (1998), also known as the Crépon, Duguet and Mairesse *(CDM)* model. This model has provided an empirical and conceptual framework to investigate the link between R&D investments, inventions and firms' economic performances that has strongly influenced the applied literature in the field of economics of innovation (Lööf et al., 2017).

The diffusion of patents as indicators of innovation, while not free from limitations (Griliches, 1990), has recently paved the way for micro-econometric studies linking this source to administrative data. This is a nontrivial exercise due to the lack of shared codification of firms' and individuals' identities in the two types of data sources. Researchers in the field have implemented name-matched procedures to overcome this drawback, with respect to both assignees' and inventors' names.

Regarding the link to firm-level data, string-matching algorithms are needed because non-standard business names are often recorded in patent documents. Major issues are related to the fact that corporations often file for patents through subsidiaries or other legal entities. This requires the tracking of business groups, as well as mergers, acquisitions and sales dynamics. Despite these difficulties, the former systematic attempt to link patents to firms is due to Hall et al. (2001). Based on this work, Balasubramanian and Sivadasan (2011) developed a dataset linking patents to the US Economic Census information. Leveraging administrative data, the authors provide evidence of the economic changes incurred by the firms following the patent introduction. The authors find that patenting is associated with significant changes within firms. In particular, their results show that the introduction of a patent is associated with a significant increase in firm size and with the introduction of new products. The association with increases in capital and skill intensity and firm-level productivity appears to be instead somewhat weaker.

More recently, Graham et al. (2018) have proposed an improved approach to matching patents and US Economic Census data by matching not only assignees' names to firms but also inventors' names to firms' employees. Hence patents are linked to two main sources of administrative data, i.e. the Longitudinal Business Database and the LEHD Employment History Files. This approach allows not only for sharper matching, reducing the occurrence of false positives, but also for the exploitation of the rich set of information contained in matched employer-employee databases.

An interesting finding reported by the authors concerns the association of firms' patenting activity with job creation rates. Their results show that patenting firms create more jobs than non-patenting firms for all age classes except among the youngest firms (those that are 1 year old). By contrast, non-patenting firms on average destroy more jobs than do patenting firms across almost all age classes, with the youngest non-patenting firms destroying the most jobs. They show that differences in the job destruction margin play a particularly important role in explaining the relative higher net growth rates of patenting firms. On average, they find that job creation rates for patenting firms exceed those for non-patenting firms by less than 1 growth point.

While R&D and patent statistics are widely used as input and output measures of innovation, respectively, the evolutionary literature grounded in Schumpeter's works also focuses on entrepreneurs as carriers of innovation in the market and, hence, on entrepreneurship and new firm formation as a measure of innovation (Schumpeter, 1911, 1942). The study of entrepreneurship has largely benefited from access to administrative data in recent years. The use of business register information and matched employer-employee datasets allows for insightful analyses linking start-up creation to demographic dynamics. The work by Azoulay et al. (2020), for example, tackles the founder's age issue, showing that, despite the widespread belief that the most successful start-ups are created by young people, in the United States the average age of the top 1000 fastest-growing new ventures' founders is about 45 years. The authors show that age predicts success, showing that the highest success rates in entrepreneurship come from founders in middle age and beyond, and that previous experience plays a key role in this respect.

Azoulay et al. (2022) explore the association between entrepreneurship and migration dynamics. Based again on US administrative data, they show that immigrants act more as 'job creators' than 'job takers' and that foreign-born founders play outsized roles in US high-growth entrepreneurship. This evidence about immigrants as agents of entrepreneurship speaks for their role in expanding labour demand relative to labour supply, contributing to balancing the debate on the interplay between migration flows and local labour market dynamics.

Moreover, Akcigit and Goldschlag (2023) use US administrative data and other data sources to investigate the career paths of inventors. Their findings indicate that (i) inventors are becoming increasingly concentrated in large established companies, are less likely to work for start-ups and are less inclined to become entrepreneurs, and (ii) when an inventor is employed by an incumbent firm rather than a young company, their earnings increase by 12.6 while their innovative output decreases by

6–11. Additionally, the authors demonstrate that these trends remain consistent and are not influenced by life cycle or occupational composition effects.

Most of the studies leveraging administrative data to study the economic drivers and effects of innovation focus on the US evidence. European-wide studies are not possible due to the decentralized maintenance of administrative information. This kind of empirical study may only be run on a single-country basis. Recent interesting evidence is brought forward by Akcigit et al. (2023), who use Italian matched employer-employee data, combined with social security and patent information, to study the interplay between political connections, industrial dynamics and innovation in Italian firms. They find that market leaders are much more likely to be politically connected but much less likely to innovate. Moreover, political connections relate to a higher rate of survival, as well as growth in employment and revenues, but not to productivity.

Overall, the lack of integrated data sources across European countries makes the implementation of large-scale empirical studies unfeasible, calling for significant collective efforts to improve the *status quo*.

5 Assessing the Causal Impact of EU R&D Grants on Firm Outcomes

Measuring the causal impact of public policies is a fundamental goal of modern economics, as it provides critical guidance on whether taxpayers' money is being used efficiently. In the domain of innovation policy, an important concern is the extent to which government interventions—such as R&D grants and subsidies—lead to improvements in firm outcomes, including innovation capacity and productivity, that would not have occurred in the absence of such support (Hünermund & Czarnitzki, 2019).

Addressing this question requires not only appropriate econometric techniques but also access to high-quality and timely data. In particular, the availability of comprehensive administrative sources covering the full population of firms that applied for public funding is indispensable.[12] These typically include information on applicant firms (e.g. VAT identifiers), the outcome of the funding decision, the amount of monetary funds disbursed and information on the selection process implemented by funding agencies, such as assessment scores and rankings assigned by evaluation panels in the context of competitive funding. Firm identifiers can be leveraged to link these administrative data with external sources reporting financial characteristics (e.g. balance sheets), as well as innovation outcomes (e.g. patents) measured both before and after funding.

[12] This makes it possible to compare firms that actively sought public funding, thereby ensuring a more credible counterfactual. In contrast, comparing funded firms to observationally similar firms that did not apply may lead to biased estimates, as the decision to apply is itself non-random and potentially correlated with unobserved characteristics.

For the researcher, an ideal setting to credibly infer the causal impact of public policies on firm outcomes would involve the random allocation of public R&D support by the funding institution. Such randomization would ensure that treated and non-treated firms are, on average, identical in all respects except for the receipt of the grant, thus enabling clean identification of causal impacts and credible evidence on what works for policymakers. However, the implementation of randomized controlled trials (RCTs) in the context of innovation policy is rarely feasible, due to a range of ethical, practical and political constraints (see Bravo-Biosca, 2020; Dalziel, 2018).

Funding institutions may face pressure to allocate resources based on merit or potential impact, and policymakers may be reluctant to exclude promising firms from support purely for experimental purposes. Hence, the main challenge lies in inferring the causal effects of public R&D support by comparing recipient firms with non-recipient firms, where these two groups are likely to differ in both observable and unobservable characteristics correlated to the outcomes of interest (i.e. propensity to innovate, firm size, etc.). In such cases, treatment assignment is endogenous, and econometric models that fail to adequately address this endogeneity are likely to yield biased and inconsistent estimates of the policy's impact.

In the light of these challenges, the economics literature has increasingly adopted Regression Discontinuity Designs (RDD) as one of the most credible alternatives to experimental evaluation (Lee & Lemieux, 2010). In the context of R&D grants, RDD leverage the institutional design of grant assignment mechanisms: firm-submitted projects are evaluated and scored by independent experts, and ranked accordingly. This ranking is then used to award funding to the highest-ranked proposals until the available budget is exhausted. This process creates a discrete cutoff based on the ranking, which can be exploited to compare firms just above and just below the funding threshold, relying on the assumption that treatment assignment near the threshold is 'as good as random'. Firms close to the cutoff are expected to be comparable in both their observable and unobservable characteristics, allowing any post-treatment differences in outcomes to be interpreted as the causal effect of receiving the grant.

To estimate the causal impact of receiving an R&D grant, in a *sharp RDD* setting[13] an equation of the following form is usually estimated (via standard OLS):

$$Y_{ic}^{\texttt{Post}} = \alpha + \beta\, \texttt{Grant}_{ic} + f(\texttt{Rank}_{ic}) + \delta_c + \varepsilon_{ic} \tag{32.8}$$

where $-r \leq \texttt{Rank}_{ic} \leq r$.

In equation (8), $Y_{ic}^{\texttt{Post}}$ denotes the post-treatment outcome for firm i applying to the "calls for proposals" c, $\texttt{Rank}_{ic}$ is the rank assigned by the evaluation panel to firm i in call c centred at the funding threshold, and $\texttt{Grant}_{ic}$ is an indicator variable

[13] Depending on the nature of the treatment assignment mechanism, the RDD can be *sharp* or *fuzzy*. While *sharp RDD* assumes perfect compliance with the threshold rule—all firms above the cutoff receive treatment and none below do—*fuzzy RDD* allows for imperfect compliance, acknowledging that some firms just below the threshold may receive funding and some just above may not. In such cases, the discontinuity in the *probability of treatment*, rather than in the treatment itself, is used to estimate a local average treatment effect (LATE) of receiving the grant.

equal to 1 if firm i is awarded the *R&D* grant or not (i.e. $\texttt{Rank}_{ic} > 0$). The function $f(\texttt{Rank}_{ic})$ is a polynomial control for the centred rank.

The regression includes also "call for proposals" fixed effects (δ_c), which restrict comparisons to infra-marginal firms (just above and below the funding cutoff) within the same competition controlling for time- and topic-specific factors. The parameter r represents the bandwidth around the cutoff (often chosen according to the criteria proposed by Calonico et al., 2020 to ensure inference-optimal bandwidth choices), and ε_{ic} is the usual idiosyncratic error term. Standard errors are robust and clustered at the 'call for proposals' level to account for potential serial correlation. In this setting, $\hat{\beta}$ identifies the local average treatment effect (LATE) for firms at the cutoff under perfect compliance and represents the causal effect of the policy.[14]

In this vein, drawing on a small ensemble of studies using RDD to estimate the casual effects of R&D grants in individual countries or regions (see, e.g., Bronzini & Iachini, 2014; Howell, 2017), recent contributions have begun to shed light on the effectiveness of R&D support channeled through the European Framework Programme[15] (Katay et al., 2025; Mitra & Niakaros, 2023; Santoleri et al., 2024). These studies leverage privileged access to administrative data from CORDA,[16] which contains application-level records from European R&D funding initiatives,[17] linked to firm-level information on financial and non-financial characteristics (e.g. Orbis), as well as innovation outcomes measured using major intellectual property databases (e.g. PATSTAT, Orbis IP).

This represents a substantial improvement over earlier studies that were based exclusively on data on beneficiaries, which limited the possibility to infer causal effects due to the absence of a suitable comparison group (see, e.g., Bayona-Sáez & García-Marco, 2010, Szücs, 2018).

The first relevant study, Santoleri et al. (2024), investigates a specific Horizon 2020 (FP8) R&D grant initiative—the *SME Instrument*—which specifically targeted individual small and medium-sized enterprises (SMEs) offering competitive grants

[14] In contrast, with a *fuzzy* RDD, a two-stage least squares (2SLS) approach should be used, where the cutoff indicator serves as an instrumental variable (IV) for actual grant receipt, thus identifying the LATE for compliers near the threshold.

[15] The European Framework Programme (FP) is the European Union's main instrument for funding research and innovation. In recent years, it has increasingly emphasized support for business R&D, including through competitive grant schemes targeting small and medium-sized enterprises (SMEs). Horizon 2020 (FP8), which ran from 2014 to 2020, had a budget of around €80 billion, while its successor, Horizon Europe (FP9), allocates over €95 billion for the 2021–2027 period.

[16] CORDA (COmmon Research DAta Warehouse) is a database managed by the European Commission Directorate-General for Research and Innovation containing information on applicants to the different Framework Programmes.

[17] This was made possible through a structured partnership, started in 2020, between the European Commission's DG Research & Innovation (Unit G.2) and the Competence Centre on Microeconomic Evaluation (CC-ME) at the Joint Research Centre of the European Commission. The Competence Centre on Microeconomic Evaluation (CC-ME) is a European Commission initiative that supports evidence-based policy by conducting counterfactual micro-econometric analyses with the aim to provide causal evidence on which policy measures work in different areas including social, industrial and environmental policies.

of up to €2.5 million. Using a sharp RDD, the paper documents that grants induce sizable effects over a number of firm-level outcomes. Firms that win grants exhibit significantly greater innovation output—measured by both the number and citation-weighted quality of patents—as well as increased investment, higher likelihood of receiving private equity or venture capital, faster growth in assets and employment and lower exit rates (see Fig. 32.5).

Importantly, the paper investigates the underlying mechanisms and provides evidence that the observed effects are primarily driven by funding, rather than certification (Lerner, 2000). To isolate the role of certification, the authors exploit the assignment of the *'Seal of Excellence'* (SOE), a quality label given to highly ranked but unfunded proposals. Results show that SOE recipients do not experience significant performance improvements relative to other non-funded non-SOE firms, implying that the grant's impact is not due to signalling alone. Instead, the findings support the idea that public funding relaxes financial frictions, particularly for smaller, younger firms and those in financially vulnerable sectors, enabling them to undertake high-risk innovation projects they would otherwise forgo.

The second contribution, Katay et al. (2025), analyses the FP7 *Cooperation Program*, which accounts for approximately two-thirds of the total FP7 budget. This programme's primary objective was to stimulate transnational collaborative research consortia, facilitating partnerships both among profit-oriented enterprises and between industry and academia within specific thematic areas. Unlike the mono-beneficiary grants studied in Santoleri et al. (2024), which targeted individual SMEs, the Cooperation Programme supported projects involving multiple partners, emphasizing collaboration between industry and academia. Results show that FP7 grants increase firms' sales and labour productivity by about 18%, whereas no effect on employment is detected. Similar to the previous study, results are consistent with the notion that public funding helps alleviate frictions—particularly financial constraints—that hinder private investment in innovation.

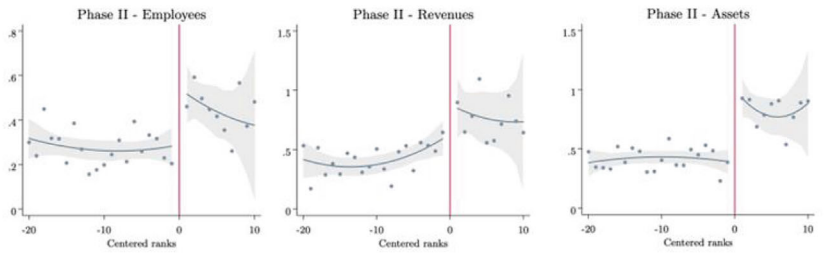


Fig. 32.5 Effects of R&D grants on firm outcomes. The figures plot firm-level outcomes (i.e. firm size in assets, employees and revenues) around the funding threshold. Circles represent rank-level means of the firm-level outcomes. Fitted lines from local polynomial regressions with a quadratic fit together with 95% confidence intervals. The figure plots the above-mentioned variables in the post-grant period, indicating that firms on the right (i.e. those receiving the grant) experience an increase in size. Source: (Santoleri et al., 2020)

Finally, a third study by Mitra and Niakaros (2023) addresses the effects of R&D subsidies by focusing on all private firms that submitted a proposal to FP8 (i.e. Horizon 2020), including grants to both mono-beneficiaries and consortia of firms. Similar to the other two papers, the authors link CORDA data with Orbis to retrieve information on firms' balance sheets. Unlike the two above-mentioned studies that rely on RDD to estimate causal effects, this paper adopts a staggered difference-in-differences (DiD) approach (Callaway & Sant'Anna, 2021).[18]

Mimicking an RDD approach, the paper limits the comparison between funded and unfunded projects that all passed the minimum quality threshold, thus discarding proposals that received low scores and that may be less comparable in both observable and unobservable terms. The study finds that, following the receipt of Horizon 2020 funding, beneficiary firms experience notable improvements in performance, with employment rising by around 20% and both total assets and revenues increasing by roughly 30% compared to similar non-funded firms.

The three papers mentioned represent a first example of how robust evidence on the impact of EU-level R&D policies requires a combination of several elements: institutional will to leverage administrative data originally collected for policy management and implementation for research purposes; efficient and generalized access to high-quality information on firms that applied for European Commission funding; the possibility to link—via standardized unique identifiers—the firm-level data with external data sources containing both firm characteristics and relevant outcomes (e.g. innovation outcomes) and, finally, suitable founding mechanisms for applying econometric techniques that allow for credible causal inference.

While significant progress has been made in recent years, much work remains to be done. This includes targeted efforts to link CORDA data across multiple Framework Programmes (e.g. from FP7 to FP9) in order to track applicants over time, improve data integration with existing firm-level databases using unique firm identifiers or reliable disambiguation procedures (as also discussed in Sect. 4) and develop a more efficient data access infrastructure—fully compliant with the existing regulations ensuring data confidentiality—through secure remote access services for research and policymaking purposes. Nonetheless, the recent advances in this area are promising.

6 Enhancing Policy with Better Data Access

This chapter shows how research can provide new policy-relevant insights using the rich administrative data available in most countries. However, several challenges remain regarding data availability, access and the day-to-day use of such data for research. Addressing these challenges could substantially strengthen the evidence

[18] This is due to the fact that Horizon 2020 has different sub-programmes with heterogeneous grant assignment procedures that make the use of an RDD challenging when considered together.

and enhance the quality of the policymaking process. Some critical aspects are highlighted here.

First, there have been many severe economic crises over the past decade and policymakers have turned to economists seeking for insights and advice. In order to be able to respond promptly to new developments, the large lags in data availability would need to be reduced. In some cases, the 'current year' available in administrative data for research is three to four years behind the calendar year. In these cases researchers can only provide general rather than specific answers to concrete questions posed by policymakers in times of crisis. As argued in Chap. 33, this is not intrinsic to administrative data nor to causal analysis, but rather to data creation and access procedures.

A second and even more serious issue is the current legal situation under which much of the data collected by the public sector must be deleted after a certain period of time (*retention period*). In many countries, statistical offices are forced by some interpretation of the regulations to delete either the data itself or at least the identifiers.

It is often said that data is the new resource of the twenty-first century (The Economist, 2017). If this is true, deleting older data means throwing away this most valuable resource. This implies a waste of valuable public resources for the original collection of data—an issue of concern in itself—but also the ability to draw lessons from past policy experiences for researchers, policymakers and taxpayers. Policies that are well intentioned may be designed which will fail to deliver because the evidence on the past has been deleted. Differentiated or extended *data retention periods* for historical, scientific and statistical purposes—implemented with appropriate safeguards—should be permitted to ensure that the substantial value of the information collected is not lost.

Third, current interpretations of data protection may not yet fully capture its intended scope. There is no doubt that data protection is an essential safeguard in a digitalized world, without which a functioning and fair economy and society cannot flourish. Privacy has increasingly become a priority to provide the highest possible level of protection and to prevent every conceivable abuse. This principle is crucial, even if not all data needs the same level of protection, and not all access to data should be treated equally.

Currently, certain interpretations of the data protection laws for example treat firm-level information on turnover, value added, employment, etc., in a similar way as information on individuals' illnesses, diseases, treatments, etc., contained in hospital records. The latter are for sure more sensitive and require much greater protection than the former. It furthermore appears controversial when a lot of corporate data must be made public due to certain laws—and is thus published on official government websites for transparency—but virtually the same data becomes confidential as soon as a researcher requires to work with it.

Recent research also underscores the importance of data access for scientific progress and evidence-based policy formulation (Nagaraj & Tranchero, 2023). Differentiated legislation exists for almost all areas of the economy and society; this should be the case also for company data, which is anonymized or pseudo-anonymized anyway once researchers work with it. In future, there will be the need

to combine much more and much more diverse types of data from many different sources to better understand the problems and challenges of an increasingly complex world. This will only be possible if the restrictions to the use of microdata by the scientific community will ensure efficient access to the data while also protecting the confidentiality of the information.

Finally, working with microdata from statistical offices and the like is sometimes difficult. Researchers—in some countries—have to travel to the location of the relevant authority or statistical office and can only use the data in dedicated safe-rooms for a limited period of time. This may be seen as a *de facto* additional barrier because it restricts researchers in their research potential. Ideas are not just generated one or two days a week during office hours in a government building but often come at any time, and it would greatly increase research productivity if these ideas could be investigated in a safe but timely manner.

The lack of facilitated access to administrative microdata means also that it takes more time to analyse and write a paper that can be published in a relevant scientific journal and/or technical report as it does to use an alternative data source. It also follows that if an important question, whether social or economic, needs to be addressed and answered quickly and with the necessary scientific rigour, the researcher is well aware of the risk to use microdata from administrative sources and the like, and hence stay away from this data. At the same time, these are often the most appropriate data because they are often the most comprehensive and detailed. The result is that society as a whole is losing out because it is not taking advantage of all the knowledge that is potentially available. In future, therefore, there should ideally be easy and practical remote access to almost all administrative data.

References

Ackerberg, D., Benkard, C. L., Berry, S., & Pakes, A. (2007). Econometric tools for analyzing market outcomes. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 6A, pp. 4171–4276). North-Holland.

Ackerberg, D. A., Caves, K., & Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica, 83*(6), 2411–2451.

Akcigit, U., & Goldschlag, N. (2023). *Where have all the "creative talents" gone? Employment dynamics of us inventors* (Technical Report). National Bureau of Economic Research.

Akcigit, U., Goldschlag, N., Baslandze, S., & Lotti, F. (2023). Connecting to power: Political connections, innovation, and firm dynamics. *Econometrica, 91*(2), 529–564.

Azoulay, P., Jones, B. F., Kim, J. D., & Miranda, J. (2020). Age and high-growth entrepreneurship. *American Economic Review: Insights, 2*(1), 65–82.

Azoulay, P., Jones, B. F., Kim, J. D., & Miranda, J. (2022). Immigration and entrepreneurship in the United States. *American Economic Review: Insights, 4*(1), 71–88.

Bajgar, M., Berlingieri, G., Calligaris, S., Criscuolo, C., & Timmis, J. (2020). *Coverage and representativeness of Orbis data.* OECD.

Balasubramanian, N., & Sivadasan, J. (2011). What happens when firms patent? New evidence from US economic census data. *The Review of Economics and Statistics, 93*(1), 126–146.

Bayona-Sáez, C., & García-Marco, T. (2010). Assessing the effectiveness of the Eureka Program. *Research Policy, 39*(10), 1375–1386.

Berry, S., Gaynor, M., & Morton, F. S. (2019). Do increasing markups matter? Lessons from empirical industrial organization. *Journal of Economic Perspectives, 33*(3), 44–68.
Bravo-Biosca, A. (2020). Experimental innovation policy. *Innovation Policy and the Economy, 20*(1), 191–232.
Bronzini, R., & Iachini, E. (2014). Are incentives for R&D effective? Evidence from a regression discontinuity approach. *American Economic Journal: Economic Policy, 6*(4), 100–134.
Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics, 225*(2), 200–230.
Calligaris, S., Criscuolo, C., & Marcolin, L. (2018). *Mark-ups in the digital era.* OECD.
Calligaris, S., Criscuolo, C., Marcolin, L., Chaves, M., Criscuolo, C., De Lyon, J., Greppi, A., & Pallanch, O. (2024). *Exploring the evolution and the state of competition in the EU*.
Calonico, S. Cattaneo, M. D., & Farrell, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal, 23*(2), 192–210.
Conlon, C., Miller, N. H., Otgon, T., & Yao, Y. (2014). *Rising markups, rising prices?* (AEA Papers and Proceedings, Vol. 113). American Economic Association (pp. 279–283). Broadway, Suite 305. Nashville, TN 37203.
Crépon, B., Duguet, E., & Mairessec, J. (1998). Research, innovation and productivity: An econometric analysis at the firm level. *Economics of Innovation and New Technology, 7*(2), 115–158.
Czarnitzki, D., Confraria, H., et al. (2025). *R&D productivity: Are ideas harder to find or does Europe suffer from a commercialization gap?*
Dalziel, M. (2018). Why are there (almost) no randomised controlled trial-based evaluations of business support programmes? *Palgrave Communications, 4*(1), 1–9.
De Loecker, J., & Warzynski, F. (2012). Markups and firm-level export status. *American Economics Review, 102*(6), 2437–2471.
De Loecker, D., Eeckhout, J., & Unger, G. (2020). The rise of market power and the macroeconomic implications. *Quarterly Journal of Economics, 135*(2), 561–644.
De Ridder, M. (2024). Market power and innovation in the intangible economy. *American Economic Review, 114*(1), 199–251.
Doraszelski, U., & Jaumandreu, J. (2013). R&D and productivity: Estimating endogenous productivity. *Review of Economic Studies, 80*(4), 1338–1383.
Draghi, M. (2024). *The future of European competitiveness*.
Duso, T., Nardotto, M., & Schiersch, A. (2025). Broadband and productivity: Structural estimates for Germany. *Journal of Economic Behavior and Organization, 237*, 107133
Graham, S. J. H., Grim, C., Islam, T., Marco, A. C., & Miranda, J. (2018). Business dynamics of innovating firms: Linking US patents with administrative data on workers and firms. *Journal of Economics & Management Strategy, 27*(3), 372–402.
Griliches, Z. (1979). Issues in assessing the contribution of research and development to productivity growth. *The Bell Journal of Economics*, 92–116.
Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature, 28*(4), 1661–1707.
Griliches, Z. (1994). Productivity, R&D, and the data constraint. *The American Economic Review, 84*(1), 1.
Griliches, Z., & Jorgenson, D. W. (1966). Sources of measured productivity change: Capital input. *The American Economic Review, 56*(1/2), 50–61.
Hall, B. H, Jaffe, A. B., & Trajtenberg, M. (2001). *The NBER patent citation data file: Lessons, insights and methodological tools.*
Hall, B. H., Jaffe, A. B., Trajtenberg, M., Griliches, Z., & Hausman, J. A. (1986). Patents and R and D: Is there a lag? *International Economic Review, 27*(2), 265–283.
Howell, S. T. (2017). Financing innovation: Evidence from R&D grants. *American Economic Review, 107*(4), 1136–1164.

Hünermund, P., & Czarnitzki, D. (2019). Innovation policy and causality. *IFO DICE Report, 17*(4), 3–6.
Jaffe, A. (1986). Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. *American Economic Review, 76*(5), 984–1001.
Jorgenson, D. W., & Griliches, Z. (1967). The explanation of productivity change. *The Review of Economic Studies*, *34* (3), 249–283.
Katay, G., Filep-Mosberger, P., & Tucci, F. (2025). *The impact of EU grants for research and innovation on firms' performance*. Magyar Nemzeti Bank (Central Bank of Hungary): Technical Report.
Le Mouel, M., & Schiersch, A. (2024). Intangible capital and productivity divergence. *Review of Income and Wealth.*
Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*(2), 281–355.
Lerner, J. (2000). The government as venture capitalist: The long-run impact of the SBIR program. *The Journal of Private Equity*, 55–78.
Letta, E. (2024). *Much more than a market*.
Levinsohn, J., & Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *Review of Economic Studies, 70*(2), 317–341.
Lööf, H., Mairesse, J., & Mohnen, P. (2017). CDM 20 years after. *Economics of Innovation and New Technology, 26*(1–2), 1–5.
Mansfield, E. (1981). Composition of R and D expenditures: Relationship to size of firm, concentration, and innovative output. *The Review of Economics and Statistics*, 610–615.
Marschak, J., & Andrews, W. H. (1944). Random simultaneous equations and the theory of production. *Econometrica, 12*(3/4), 143–205.
Mitra, A., & Niakaros, K. (2023). *The Horizon effect: A counterfactual analysis of EU research & innovation grants number 2023/08* (Research and innovation paper series, 2023).
Monopolkommission. (2022). *Hauptgutachten XXIV: Wettbewerb 2022.*
Monopolkommission. (2024). *Hauptgutachten XXV: Wettbewerb 2024.*
Nagaraj, A., & Tranchero, M. (2023). *How does data access shape science? Evidence from the impact of US census's research data centers on economics research* (Technical Report). National Bureau of Economic Research.
Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica, 69*(2), 307–342.
Olley, S., & Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica, 64*(6), 1263–1297.
Pakes, A., & Griliches, Z. (1984). *Patents and R&D at the firm level: A first look. R&D, patents, and productivity* (pp. 55–72). University of Chicago Press.
Pavitt, K. (1982). R&D, patenting and innovative activities: A statistical exploration. *Research Policy, 11*(1), 33–51.
Santoleri, P., Mina, A., Di Minin, A., & Martelli, I. (2020). *The causal effects of R&D grants: Evidence from a regression discontinuity* (Technical Report, LEM Working Paper Series).
Santoleri, P., Mina, A., Di Minin, A., & Martelli, I. (2024). The causal effects of R&D grants: Evidence from a regression discontinuity. *Review of Economics and Statistics, 106*(6), 1495–1510.
Schumpeter, J. (1911). *The theory of economic development (1934 translation).* Transaction Books.
Schumpeter, J. (1942). *A capital, socialism, and democracy*. Harper.
Slade, M. E. (2004). Market power and joint dominance in UK brewing. *The Journal of Industrial Economics, 52*(1), 133–163.
Solow, R. M. (1957). Technical change and the aggregate production function. *The Review of Economics and Statistics, 39*(3), 312–320.
Szücs, F. (2018). Research subsidies, industry-university cooperation and innovation. *Research Policy, 47*(7), 1256–1266.

The Economist. (2017). The world's most valuable resource is no longer oil, but data. *The Economist (Leaders)*. Available at: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Van Beveren, I. (2012). Total factor productivity estimation: A practical review. *Journal of Economic Surveys, 26*(1), 98–128.

Stefano Verzillo is a Senior Economist currently working at the European Commission's DG Joint Research Centre (JRC), Competence Centre on Microeconomic Evaluation (CC-ME). He is also Adjunct Professor of Statistical Modeling, Data Mining and Machine Learning at University of Milano-Bicocca and IZA@LISER Research Fellow. He holds a Phd from University of Milan (2013). His research focuses on applied microeconometrics covering a variety of fields in Education and Health Economics, Economic Inequality and Programme Evaluation. His work relies extensively on administrative data and quasi-experimental methods to study how reforms, incentives, and public policies affect individuals, households, firms, and research organizations.

Tomaso Duso is Professor of Empirical Industrial Economics at the Technische Universit t (TU) Berlin and Head of the Firms and Markets Department at the German Institute for Economic Research (DIW Berlin). He is also Chairman of the German Monopolies Commission, a member of the European Commission's Economic Advisory Group on Competition Policy (EAGCP) and a Research Fellow of CEPR and CESIfo. Tomaso holds a PhD in Economics from the Humboldt University of Berlin. His research interests are in applied econometrics in the fields of industrial organization, competition policy, regulation and management. He has advised several public institutions on competition policy issues, including several DGs of the European Commission, the UK Competition Commission and Competition and Markets Authority, the Dutch Authority for Consumers and Markets, the OECD, and the European Bank for Reconstruction and Development.

Francesco Quatraro is Full Professor of Economics at the University of Torino and Fellow at the Collegio Carlo Alberto. He is co-editor of the journal Spatial Economic Analysis and the Italian Economic Journal. At the University of Torino he is Director of the ICxT Innovation Center, and Vice-Director of the PhD Programme in Innovation for the Circular Economy. His research focuses on the economic determinants and impacts of technological innovation at the firm and regional level, with a specific focus on the interplay between new technological challenges, sustainable development and labour market dynamics. He is an applied economist with sound experience in the implementation of sophisticated econometric techniques for the analysis of firms' and regions' economic and innovative performances, and for the evaluation of the impact of policies through counterfactual analysis.

Pietro Santoleri is an Economic Analyst at the European Commission's Joint Research Centre (JRC) in Seville and an affiliated Researcher at Sant'Anna School of Advanced Studies. He holds a PhD in Economics jointly from Paris I Pantheon-Sorbonne and Sant'Anna, obtained in 2019. He has previously worked for the United Nations (Chile) and STATEC (Luxembourg) and served as consultant for the Organisation for Economic Co-operation and Development (OECD, Paris). His primary fields are firm dynamics, economics of innovation and science of science, with an emphasis on the evaluation of research and innovation policies.

Alexander Schiersch is a Senior Economist at the German Institute for Economic Research (DIW Berlin) and member of the Scientific Advisory Group (Wissenschaftlicher Beraterkreis - WBK), which advises the Research Data Centres of the German Statistical Offices. His research focuses on the role of intangible capital, technological progress and public goods in shaping firm dynamics and competitiveness. More recently, he has focused on the factors influencing corporate carbon

dioxide emissions. In addition to publishing in relevant economic journals, he is actively involved in policy-oriented projects.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 33
Current Challenges and Opportunities in Data-Driven Learning



Fabio Berton, Paolo Paruolo, and Ian Vollbracht

Abstract This chapter discusses some of the current challenges and opportunities in data-driven learning, including the role played by Artificial Intelligence, data protection techniques and human capital. It argues that computational systems can accelerate and support (but not replace) humans in data-driven learning, that advances in data protection can improve privacy and access to microdata for research, and that human capital is a key factor for data-driven learning, both in the production of causal evidence and in its communication and use in society. Finally, the chapter discusses the association between data-driven learning and trust in the public sector in democracies, which further reinforces interest in policy measures that support data-driven learning.

1 Introduction

The traditional multi-year policy cycle[1] has been called into question by several phenomena that unfolded in the first half of the 2020s. One of them was the COVID-19 pandemic, which revealed the need for rapid policy decisions coupled with swift

[1] In political science, the term 'policy cycle' is understood as a tool for analysing the stages of a policy's lifecycle. It is called the policy cycle because the final stage (evaluation) often leads back to the first stage (problem definition), thereby restarting the cycle.

The views expressed in this paper are those of the authors and do not necessarily reflect the opinion of the institutions of affiliation. The authors thankfully acknowledge useful comments from Stefano Verzillo on a previous version of this chapter.

F. Berton (✉) · P. Paruolo · I. Vollbracht
European Commission Joint Research Centre, Ispra, Italy
e-mail: fabio.berton@ec.europa.eu

P. Paruolo
e-mail: paolo.paruolo@ec.europa.eu

I. Vollbracht
e-mail: ian.vollbracht@ec.europa.eu

© The European Union 2026

F. Berton and P. Paruolo (eds.), *Data-Driven Learning in the EU*, Professional Practice in Governance and Public Organizations, https://doi.org/10.1007/978-3-031-99032-8_33

evaluations about their effectiveness (Berta et al., 2020; Bertone et al., 2020; Cerqueti et al., 2022). Other phenomena included the increasing relevance of trade tariffs, energy price volatility and geopolitical conflicts, which have also underscored the need for faster and more adaptive policy responses.

The same period has also witnessed the surge and diffusion of large language models (LLM)[2] (Gambacorta et al., 2025; OECD et al., 2025). They can be used, inter alia, to spread any belief, potentially even those that run contrary to what science has shown to be effective. This challenges the way science is communicated and perceived in societies (Diel et al., 2022; Himelein-Wachowiak et al., 2021). All of the aforementioned phenomena highlight the need for faster, more effective and adaptive policies, informed by timely scientific evidence.

The need for fast policy turnaround time contrasts with the (usually slower) pace of science, research and development. There are different solutions on how to improve timeliness and abundance of causal evidence to best serve policy decisions. The one proposed below requires planning the access to microdata, and investing in human capital and technology, including Artificial Intelligence (AI).

Causal evidence of policy-effectiveness involves data-driven learning, defined as the process of learning (if and) how policies work and for whom. The use of relevant and reliable microdata emerges as a crucial ingredient for causal evidence-making. This chapter discusses the lessons derived from previous chapters, with a view to outlining what institutions could do to improve upon the status quo with a focus on the European Union (EU).

Several messages emerge. In spite of the shared EU legislative framework, the book uncovers 27 different combinations of technical and procedural approaches to microdata access. There are several good practices to imitate, even for those countries in which rules governing access are clear and well organised. Moreover, there is much more knowledge potential in administrative microdata than is currently being made available and exploited by researchers. These observations call for the identification of challenges and opportunities—and thus areas for intervention—to enhance access to administrative microdata for research purposes.

Data access is not an end in itself, but rather the means to contribute to more reliable evidence-based decision-making. Policy effectiveness is best evaluated with causal methods, i.e. appropriately applying counterfactual impact evaluation methods to reliable (micro)data; the same applies to any postulated causal relationship. Experience shows that a number of elements need to be in place first, though.

The first element is that the relevant administrative data must exist, be available and that it must be accessible to researchers, under conditions that are fully compliant with data protection rules, such as the General Data Protection Regulation (GDPR). Privacy and confidentiality of data are a pivotal concern in the data-driven learning

[2] Artificial Intelligence (AI) is the capability of computational systems to perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception and decision-making. Machine learning (ML) is a field of study in AI concerned with the development and study of statistical algorithms that can learn from data and generalise to unseen data. A large language model (LLM) is a language model trained with self-supervised ML on a vast amount of text, designed for natural language processing tasks, especially language generation.

process; advances in data protection can improve privacy and access to microdata for research at the same time, in line with the analysis by Abraham et al. (2017). Moreover, microdata access is sometimes legally foreseen but difficult to obtain in practice: access needs to be there in theory and in practice.

The second element is that there must be the possibility to construct valid identification strategies that align with the policy evaluation question at hand. Sometimes researchers let the data select the research question, i.e. they frame research around what is measurable or identifiable rather than what is substantively important. This may be acceptable in exploratory settings, but in an applied policy context, the research question is typically given exogenously by the policy design. The evaluation must, therefore, be methodologically tailored to the question, not vice versa. AI plays a role in accelerating, supporting but not replacing humans in the evaluation design.[3]

The third element is that credible counterfactual impact evaluations can fail to generate policy learning if their findings are ignored, misunderstood or judged as being politically inconvenient. Even where data and methods are aligned, an evaluation's societal value depends on whether the results are taken seriously by decision-makers for future policy revisions. This willingness cannot always be assessed ex ante, but it is shaped by the political and institutional contexts in which the evaluation is embedded.

These three elements are not independent; weakness in one area can compromise the overall potential for policy learning. They are discussed throughout this chapter by means of a production function and a diffusion function. The roles of AI and human capital are discussed in these areas. The conclusion is that a lot more could be done to improve the institutional setting, so as to allow more value added to be produced in evidence-making, data-driven learning and causal evidence communication.

The rest of the chapter is organised as follows: Sect. 2 presents a production function and diffusion function narrative; Sect. 3 discusses developments in the relevant European legislation; Sect. 4 overviews how interactions between researchers and data can be organised so as to ease access while simultaneously improving safeguards for data protection; Sect. 5 describes how computational systems both can and cannot be used to help research; Sect. 6 discusses human capital; and Sect. 7 concludes.

2 A Production Function Narrative

This section introduces a production function narrative that will be used throughout this chapter.

Let x indicate causal evidence, i.e. evidence on the existence and form of a causal relationship between a policy intervention and some outcome of interest.

[3] See Jacobzone et al. (2025) for an overview of different uses of AI in this process. Chodacki et al. (2024) describe how supervised AI can contribute to data repository management.

Let also y indicate public value associated with better policies (because they are more targeted, less costly, more effective or efficient, etc.). Consider the following production function for causal evidence x and public value y:

$$x = f(D, L, K, A), \tag{33.1}$$

$$y = g(x, L, A), \tag{33.2}$$

where D is data, L and K are labour and capital, A is technology, and f, g are functions.[4] The function f can be understood as a semi-endogenous growth production function, as described in Jones (2022); g is akin to an innovation diffusion function as in Guidolin and Manfredi (2023).

Causal evidence x is produced via the production function f, with factors of production consisting of amount of data accessed D, labour L (the time of researchers, data-holding institutions' employees, policymakers), capital K (hardware, institutions) and technology A (ideas, theoretical and empirical models, hardware and software advancements). Once causal evidence x is produced, it is communicated and used via the function g that transforms causal evidence into public value y.

Policy interventions that aim to improve access to administrative microdata should focus on increasing public value y. Assuming that x is an increasing function of D and that y is an increasing function of x (i.e. that more data would bring more policy-relevant knowledge via research),[5] the first goal could be to increase the information D that is made available to produce causal evidence x.[6] Increasing D can be achieved e.g. by making more data both accessible and linkable across administrative sources.[7] Aspects related to data availability are discussed in Sects. 3 and 4 below.

Technology A includes machine learning and AI, and it plays a role in both the production function f and the diffusion function g, while capital K is assumed to influence only the production function f. Aspects related to technological progress A (and possibly capital K) are discussed in Sect. 5 below.

Labour L enters both functions f and g, with the expectation that more labour L increases x and y. Possible policy interventions might aim to improve the quantity of labour and human capital L and technology A in f and g, while safeguarding privacy. Aspects related to labour and human capital are discussed in Sect. 6 below.

There are several promising areas in which policymakers may intervene, namely what Chap. 29 calls interactions among data users and between users and data, the use of machine learning and AI in research, the legal framework and the efficiency

[4] One expects both functions f, g to have positive derivatives with respect to all of their arguments, see below.

[5] This is not trivial: additional data can deliberately be used or created to misinform or to exhaust the attention span of the audience.

[6] Note that administrative data are collected for the functioning of the public administration that hold the data, and any request to collect more variables would result in increased costs and administrative burdens that the data holder would not necessarily be willing to incur.

[7] Administrative data are often pre-existing somewhere.

of the causal evidence diffusion function g, where, for instance, certification of the scientific standing of causal evidence could be one way in which to support the diffusion of credible causal evidence. These issues are discussed in the concluding Sect. 7.

3 Developments in EU Data Protection

This section discusses some legal aspects related to access to data D, which influences the knowledge production function f in Eq. (33.1).

Since its entry into force on 25 May 2018, the GDPR has provided the principal legal framework for personal data protection in the European Union. The GDPR has clarified the conditions under which such data may be lawfully processed for scientific purposes, while also imposing safeguards and accountability requirements, for researchers and public authorities working with administrative microdata. In the years following its implementation, case law and regulatory guidance have progressively shaped the operational space for data-driven research.

Several judgments and official opinions have directly addressed the intersection of data protection and scientific research. These have reinforced the principle that research is a purpose for processing personal data, including administrative microdata, that deserves special legal treatment, provided appropriate safeguards are in place. The European Court of Justice in ECJ (2018) affirmed that personal data used for research purposes falls within the GDPR's scope, but also recognised that such processing may benefit from certain exemptions or derogations, especially when it is used in the public interest.

The European Data Protection Board in EDPB (2019) offered a more detailed framework. It explicitly acknowledged that administrative microdata can be highly valuable for both research and policy evaluation. However, it emphasised the importance of transparency towards data subjects, data minimisation and proportionality, pseudonymisation or anonymisation where possible, data protection by design and by default, especially in research projects' planning stages. Researchers must ensure that data subjects receive clear, accessible information about how their data are processed, even when secondary use is involved.

Further guidance came with EC (2020), which reiterated that the processing of personal data for scientific research purposes is compatible with the GDPR, but must be accompanied by robust safeguards. In particular, it stressed the need to assess data protection risks when research involves large-scale or sensitive datasets, and it encouraged the development of institutional frameworks to ensure compliance.

Both the opinion and the decision made specific mention of administrative microdata, recognising its potential to support evidence-based policymaking, particularly in the social sciences and economics. At the same time, the need for institutional arrangements to guarantee compliance is also emphasised, such as trusted research environments and standardised data governance protocols.

These developments have both enabled and constrained the use of administrative microdata for research institutions and policymakers. On the one hand, the GDPR and subsequent guidance have provided clearer support for research, increased the credibility of research practices and have encouraged investments in secure infrastructure that facilitates lawful access. There is now greater legal certainty around the secondary use of personal data for public interest research, provided that safeguards are in place.

On the other hand, the framework imposes non-trivial compliance obligations. These include strict limitations on purpose specification, rights to access and to erasure and the need to justify any exemptions under Article 89 of the GDPR. Moreover, the burden of demonstrating compliance rests squarely on data controllers, including universities and national statistical agencies.

In sum, the post-2018 evolution of EU data protection law reflects a growing recognition that research using administrative microdata can deliver significant public value, but only where pursued under conditions that respect individual rights, minimise risk and ensure accountability. The GDPR provides the framework for this balance; its implementation depends upon the capacity and commitment of institutions to integrate legal, ethical and technical safeguards into the full research lifecycle.

4 Evolving Data-Access Methods

This section discusses infrastructure and institutional arrangements that allow researchers to access and to work with administrative datasets. These arrangements are critical in shaping what types of analysis are possible, under what conditions and by whom; this influences causal evidence-making x via availability of data D in the knowledge production function f in Eq. (33.1).

A first-generation model of access involved the transfer of a (usually pseudonymised) dataset to individual researchers, typically following strict confidentiality agreements. While this model allowed for significant flexibility in analysis, it also involved a complete data transfers, raising data protection concerns and limiting scalability (Desai et al., 2016).

Several alternative models have emerged since then. These models reflect a broader shift from moving data to moving people and/or moving code, where analytical logic is brought to the data rather than the other way around. Each model represents a different balance between accessibility, control and institutional trust, and their suitability may vary depending on the specific context, dataset and environment.

On-site safe environments. In this arrangement, researchers are permitted to access data within the data holder's premises on secure servers, but no dataset ever leaves the data-holding institution. This approach eliminates external transfers and enhances the ease of compliance with data protection rules. However, it also introduces practical limitations, such as the need for researchers to be physically present in a safe room and this complicates collaboration and imposes logistical burdens on

project timelines and monetary costs for data users. This model is used by Eurostat and by many national statistical institutes.[8]

Embedded researchers. A second approach, exemplified by the 'surrogate access' at the Bank of Portugal,[9] involves pairing external researchers with internal analysts from the data-holding institution. The internal partner executes the code developed jointly with the external researcher, thereby ensuring that the raw data remains unseen and on-site.

This method can balance data protection with analytical flexibility. Beyond technical access, it fosters institutional learning and builds long-term professional collaborations, as internal staff gain greater familiarity with the analytical uses of their data—potentially making future collaborations more efficient and more open.

However, one risk related to this approach is the possibility of informal gatekeeping and of favouring specific research topics by the embedded researcher, in case her/his interests are not fully aligned with the ones of the data-holding institution; attention needs to be paid to designing this data-access model in a way that harmonises institutional and personal interests, possibly with transparent and public rules and recurrent scrutiny.

Secure remote interfaces. A third approach uses a technical interface or API that allows researchers to interact with the data—e.g. running code, requesting results—remotely without any possibility to download them or to export computations before a privacy check by the data holders. This approach exists in Finland, France, the Netherlands and Sweden, to mention a few.

In some cases, the researcher is not even allowed to visualise the data, but only aggregate information (e.g. estimates) extracted therefrom. For instance, the Norwegian Centre for Research Data (NSD) and Statistics Norway have pioneered versions of this model, offering remote platforms with secure virtual desktops and integrated helpdesk support.[10] Functionally, it is similar to the embedded researcher model, but it replaces the human intermediary with a digital layer. This model improves scalability and remote access while still preserving high standards of confidentiality and security.

Data spaces. A fourth approach is that of data spaces. A data space is a decentralised system that enables organisations to share, access and reuse data (Bacco et al., 2024). Instead of concentrating all information in a single database, a data space links independent systems through common standards, protocols and governance rules. This design allows participants—be they public institutions, companies or individuals—to collaborate, while also maintaining control over their own data. Chapter 9 on Estonia provides an example of a data-space approach, while the European Strategy for Data seeks to build a single market for data across the continent (Farrell et al., 2023).

Every data space rests on three main components. The infrastructure layer provides the trusted digital environment in which data can be stored, transferred and processed

[8] https://ec.europa.eu/eurostat/web/microdata.

[9] See Chap. 23.

[10] https://www.ssb.no/en/data-til-forskning/utlan-av-data-til-forskere.

safely. The governance layer sets the rules of participation; it determines who can access or share specific data, under what conditions and with what guarantees of transparency and accountability. Finally, the federation and interoperability layer ensures that different systems and organisations can connect and exchange information, even if they operate under distinct technologies or management structures.

Together, these layers create an environment that balances openness with control, making it possible to cooperate efficiently without centralising ownership or decision-making. In addition to providing remote access, the data spaces and federated learning place particular emphasis on data interoperability.

Synthetic populations and differential privacy. A synthetic population is an artificially generated representation of a real population that reproduces its statistical structure, distributions and relationships among attributes without a one-to-one correspondence with individuals from the real population (Hradec et al., 2022). This approach is useful for preserving privacy while testing computer code, but it is not an adequate replacement for the analysis of the real population data.

The creation of a synthetic population typically begins with an anonymised source of microdata, such as the ones from administrative registers, and the quality of the resulting data is commonly assessed through its *utility*, defined as the degree to which the synthetic dataset reproduces the analytical properties of the original data, where a high-utility dataset should yield statistical results indistinguishable from those derived from real data. The resulting dataset can be freely shared and analysed without risk of re-identification because no synthetic record corresponds to a real statistical unit.

In terms of their analytical potential, synthetic populations represent a major innovation in privacy-preserving data sharing, where methods such as anonymisation, aggregation or suppression of sensitive variables reduce disclosure risk but still leave data vulnerable to re-identification through linkage with external sources. Synthetic populations can be openly shared without concerns of legal frameworks such as the GDPR and, most importantly, of any disclosure of individual information.

At the same time, synthetic data cannot replace the original data when generating causal evidence that can directly inform policy decisions. Because synthetic and real data differ,[11] synthetic data are best suited for safe software development, methodological testing and the preliminary exploration of analytical frameworks without exposing confidential information. Once validated, however, the analysis should be executed on the real data, within secure environments, thereby ensuring that sensitive information never leaves protected infrastructures and that the resulting evidence is not biased by potential limitations in the synthesis process.

Related advanced techniques concern differential privacy (Dwork, 2008), which introduces mathematically defined noise to real data to guarantee that the inclusion or exclusion of any individual in the source data does not materially affect the output.[12]

[11] Often in unknown aspects.

[12] Differential privacy is founded on the principle that any computation performed on a dataset inevitably discloses a modicum of information about the individual records that it contains, regardless of the protective measures employed. Massive use of a same dataset may therefore reveal

Because of this characteristic, differential privacy techniques offer privacy-protection while guaranteeing closeness to the real population, and hence represent a compelling strategy for enhancing data protection.

5 Computational Systems

This section discusses how technical progress A influences causal evidence-making x via the production function f in Eq. (33.1) and public value y via the diffusion function g in Eq. (33.2), concentrating attention on the role of AI.

The advent of modern digital computers[13] has greatly modified the way data are stored and analysed, especially with statistical techniques. Improvements in statistical learning were encapsulated in the term 'machine learning', which describes the capacity of statistical algorithms to gain knowledge by identifying patterns in raw data (Goodfellow et al., 2016). Recent advances in machine learning led to the coining of the term AI, see OECD (2024), as reported by Branchini and Vallina Acha (2025). The effective use of AI depends on context and specifically on the user's skills in prompting questions so as to gain maximum benefit from its capabilities, as well as her/his capacity to critically assess its output.

A clear distinction needs to be made here between computer-assisted decisions, where the design of the analysis and related decisions are fully controlled by a human while computers perform heavy duty calculations, and AI, where at least some of the design and decisions in the data analysis are performed by the machine, see Paruolo (2005). The reason for this distinction is that causal evidence communication (via the function g) requires the full documentation of the inference steps performed in data-driven learning; these cannot be fully communicated unless humans have control over what inference is performed on which data.[14]

The support given by computers in calculations is indispensable, and machine learning is commonly used in causal microeconometrics; examples include targeted maximum likelihood estimation (TMLE) (van der Laan & Rose, 2011), augmented

some individual-specific information. It is thus essential to quantify, monitor and constrain the total privacy loss that results from repeated analyses in order to prevent cumulative privacy degradation. Differential privacy offers a formal framework for defining and managing this cumulative loss. Achieving differential privacy requires the deliberate addition of random statistical noise to the outcomes of statistical or machine learning analyses and thus introduces some reduction in analytical accuracy. Differential privacy ensures that an individual's data have a negligible impact on the analytical outcome, effectively rendering the results as though that individual's data had not been included at all. Consequently, differentially private analyses are suited for contexts in which the objective is to infer properties of populations (as research aims to do) rather than to draw conclusions about specific individuals. See Altman et al. (2020) for more details.

[13] Modern digital computers were first introduced in the 1940s. For a history of computers and information machines, see Campbell-Kelly et al. (2023).

[14] AI may fail to exhibit genuine understanding, intentionality or the cognitive capacities required for rigorous causal reasoning, see Bender et al. (2021) and Marcus and Davis (2019).

inverse probability weighting (AIPW) (Kurz, 2022) and double/debiased machine learning (DML) (Ahrens et al., 2025).

The nature and architecture of AI present a number of foundational limitations in the context of causal inference on what works. A first limitation is that AI results are context dependent, because the framing of the question and the sequencing of wording influence the answer. A second basic limitation is the lack of reproducibility of results, related both to the lack of workflow documentation and to difference in AI's output, even with the same prompting.[15]

Moreover, AI has been shown to confidently output incorrect information, known as hallucination (Sun et al., 2024). This extends beyond factual errors, to plausible-sounding causal claims that are empirically unfounded or methodologically incoherent. These aspects violate basic standards of scientific work, which include soundness of the approach, replicability, documentation of reasoning and stability of results.

Further critical aspects of AI relate to confirmation bias and accountability (Brodeur et al., 2020). AI poses epistemic and ethical challenges that are in-built in its technology, which is inherently not transparent (black box) in a context of regulatory analysis and of evaluation of large-scale social policies, in which transparency and auditability are essential. AI risks embedding confirmation bias by reinforcing prevailing narratives, particularly when trained on literature where positive findings are overrepresented. Moreover, an accountability concern also exists: if a government report incorporates code or claims drafted by AI that turn out to be flawed, who bears responsibility?

These limitations matter. In causal inference, estimators need to be well understood and used appropriately, identification strategies must be transparent and defensible, appropriate data needs to be accessed and protected properly and the whole chain of inferences needs to be documented and subject to scrutiny and replication. AI, at least at present, does not provide any of this. This is well stated in the following quote from Candès et al. (2019).

> We do not have intelligent machines. We particularly applaud recognition that much progress is needed in terms of inferring and representing causality. This is an area where the ingredients that have made AI very successful–trillions of examples, immense computing power, and fairly narrow tasks–have limited applicability. [...] counterfactual reasoning [appears] as the ability to think causally; this is absent from the current predictive machine learning toolbox. [...] In contrast, counterfactual reasoning and imagining what is not there (yet might be) are not foreign to statistics. [...] It is critical to carefully quantify our decision making algorithms, their fairness, their real-world consequences, and their confidence and robustness in predictions. These challenges should be a clarion call for statistical thinking.

It is wrong, however, to dismiss AI entirely, at least the part supporting computer based decisions overseen by humans. In a scoping review, Branchini and Vallina Acha (2025) mapped current use of AI in public policy evaluation; they found that AI is used in different context in most phases of evaluations.

[15] See the mission of the Institute for Replication (https://i4replication.org/index.html) and Brodeur et al. (2024). AI may generate false confidence on the part of the user regarding AI's output (Aslett et al. 2024).

AI can support empirical research in various auxiliary functions, particularly for applied researchers working with administrative microdata. In particular, (i) it can help with literature synthesis and summarisation, because it can quickly identify relevant studies or similar policy evaluations by using semantic search and summarisation capabilities. Moreover, (ii) it can help with hypothesis generation, suggesting plausible causal paths or patterns for exploration, especially in early-stage project formulation.

AI can also serve to (iii) accelerate workflow, producing scaffolds for (first drafts of) code in languages like R, Stata or Python, and automating mundane tasks such as data wrangling, graphing or LaTeX formatting. Finally, it can (iv) support communication, by drafting non-technical summaries, policy briefs or visual parts that accompany formal econometric work.

These capabilities are not trivial. By reducing the fixed costs of undertaking a research project, AI may help to lower the costs of knowledge production and the potential for learning across many fields. This applies to economics and data analysis, as it does to many other disciplines.[16] Moreover, AI technology continues to advance at a rapid pace and the list of use cases is expected to grow at a similar pace.

When used judiciously by trained researchers who understand their limitations, AI can accelerate the exploratory phases of research, help structure communication, especially for policy translation, and facilitate the creation of reproducible workflows, provided that it is properly integrated into version-controlled pipelines and appropriately reviewed. This requires AI users to treat AI as an assistant rather than as an infallible oracle.

6 Human Capital

This section discusses the role of labour and human capital L in knowledge production function f in Eq. (33.1) and in the diffusion function g in Eq. (33.2).

Human capital keeps being a key ingredient for data-driven learning. It is in fact central in the design of impact evaluation (new methodologies, novel use of existing methods, innovation in data use), as well as in the communication of causal evidence and its use in future policy design. While technological progress accelerates the completion of tasks, human capital appears key for analysis design and results interpretation.

Human capital is involved in (a) innovative evaluation solutions, (b) careful overview of execution, and (c) the interpretation of the findings (function f) and the diffusion of causal evidence (function g). Finding a natural (local) randomisation event with observational, non-experimental data is essentially an art, see (a), and

[16] The view of AI as a driver of democracy in research is not unchallenged, however. Ahmed and Wahed (2020) provide evidence that research in AI is subject to an increasing divide, due to unequal access to computing power between large firms and elite universities on the one side, and non-elite universities on the other. This may induce the same divide in the application fields of AI.

thus a human capital game as shown, e.g. in Cipollone and Rosolia (2007) or Pinotti (2017). Evidence on (b) can be found in the peer review system of scientific publications; the interpretation of what scientific results really mean is also not obvious, and this requires the best use of human capital, see (c).

On the use-side, Jacobzone et al. (2025) state that data-driven learning

> ...involve[s] cultural changes, including developing environments where identification of unsuccessful programmes is seen as a step in the right direction rather than a setback, where evaluation is utilised throughout policy processes, and where learning-by-doing is actively encouraged.

All of this requires a better understanding of causal evidence by humans and their collective willingness to act on it, i.e. that humans stay 'in the loop', see Branchini and Vallina Acha (2025). Researchers and institutions need to ensure that tools like AI are used in ways that complement—rather than undermine—the hard-won rigour of causal analysis in an age of disinformation and media polarisation in the public discourse.

7 Conclusions

Results from the application of proper scientific methods to data are useful as long as they are credible, can be communicated transparently and can be understood as causal effects. Low appetite for reality checks (i.e. based on science) and transparency are issues for democracies; recent years have seen a decline of trust in public institutions (Valgarðsson et al., 2025; Almeida, 2026). In this context, data-driven learning can contribute towards improving trust in the public sector, and strengthen democracy.

There are several risks that institutions face, especially those responsible for regulating the access to data and for the use of the resulting causal evidence to shape policy interventions. The first risk is to regulate data access for research purposes without taking best practices into account, as described in this book. Lack of uniform guidance on regulations may lead to overly restrictive interpretations that hinder data access beyond the legislators' intentions.

This risk can be avoided and turned into an opportunity by supporting the creation of EU-wide or National simple public rules and standards on how to access administrative microdata for research purposes. Supporting measures could include funding to create appropriate institutional settings and seed grants for researchers to use these new (or reformed) facilities. The goal is to make administrative microdata easily accessible by default, and to incentivise their use by researchers for data-driven learning.

Supporting and fostering systematic use of administrative microdata implies planning an initial infrastructure investment. However, as discussed in Chap. 1, this comes with an overall savings in operating costs, and better causal evidence. Moreover, and even more relevant, this investment can also lower reporting burden if associated with automated reporting, i.e. shifting data-related reporting from humans

to machines. This can create large savings and help implement the Commission priority to streamline reporting obligations and reducing administrative burden, see EC (2023).

A second risk is to take the shortcut of automating causal evidence on policy using technology and AI-based solutions. As discussed above, AI needs to be integrated to support causal evidence creation without renouncing human supervision on the design of analysis and interpretations of findings.[17]

Humans are essentials to supervise the production function of learning what works. Investment in human capital, as argued above, remains pivotal to design analysis and interpret results. Again, this risk can be turned into an opportunity by funding fora for human interactions on what works (production function f) and causal evidence diffusion (diffusion function g).

On the side of causal evidence diffusion, certification of scientific evidence can help to separate scientific claims from unscientific ones, which can better inform both the public and policymakers. This may benefit from the work of appropriate National or EU institutions. These are not in place in all EU Member States at the moment, and EU institutions could provide this support on the basis of the subsidiarity principle.

References

Abraham, K. G., et al. (2017). *The promise of evidence-based policymaking: Report of the commission on evidence-based policymaking*. US Census Bureau. Retrieved 1 Feb, 2026, from https://www2.census.gov/adrm/fesac/2017-12-15/Abraham-CEP-final-report.pdf

Ahmed, N., & Wahed, M. (2020). *The de-democratisation of AI: Deep learning and the compute divide in artificial intelligence research* (Technical report, Arxiv, Cornell University). https://doi.org/10.48550/arXiv.2010.15581

Ahrens, A., Chernozhukov, V., Hansen, C., Kozbur, D., Schaffer, M., & Wiemann, T. (2025). *An introduction to double/debiased machine learning*. https://doi.org/10.48550/arXiv.2504.08324

Almeida, M. (Ed.) (2026). *Scoping report: Future challenges to democracy* (Technical report JRC139315). Publications Office of the European Union. https://doi.org/10.2760/8819294

Altman, M., K. Nissim, S. Vadhan, & A. Wood (2020). *Designing access with differential privacy* (pp. 173–242). Abdul Latif Jameel Poverty Action Lab (J-PAL). Retrieved 1 Feb, 2026, from https://admindatahandbook.mit.edu/book/v1.0/index.html

Aslett, K., Sanderson, Z., Godel, W., Persily, N., Nagler, J., & Tucker, J. (2024). Online searches to evaluate misinformation can increase its perceived veracity. *Nature, 625*, 548–556. https://doi.org/10.1038/s41586-023-06883-y

Bacco, M., Kocian, A., Chessa, S., Crivello, A., & Barsocchi, P. (2024). What are data spaces? Systematic survey and future outlook. *Data in Brief, 57*, Article 110969. https://doi.org/10.1016/j.dib.2024.110969

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, New York, NY, USA, pp. 610–623. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

[17] AI companies are recognising the business opportunity to serve scientists by creating specific tools for researchers, see, e.g., https://prism.openai.com/.

Berta, P., Paruolo, P., Verzillo, S., & Lovaglio, P. G. (2020). A bivariate prediction approach for adapting the health care system response to the spread of COVID-19. *Plos one, 15*(10), 1–14. https://doi.org/10.1371/journal.pone.0240150

Bertone, E., Juncal, M. J. L., Umeno, R. K. P., Peixoto, D. A., Nguyen, K., & Sahin, O. (2020). Effectiveness of the early response to Covid-19: Data analysis and modelling. *Systems, 8*(2), 1–17. https://doi.org/10.3390/systems8020021

Branchini, B., & Vallina Acha, B. (2025). Use and application of artificial intelligence in public policy evaluation. A scoping review. *Journal of Policy Evaluation, 1*, 114–133. https://doi.org/10.4995/jpeval.2025.23837

Brodeur, A., Cook, N., & Heyes, A. (2020, November). Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review 110*(11), 3634–60. https://doi.org/10.1257/aer.20190687

Brodeur, A., Mikola, D., & Cook, N. (2024). *Mass reproducibility and replicability: A new hope* (Technical Report 16912, IZA Discussion Paper). Archive: https://www.iza.org/publications/dp16912

Campbell-Kelly, M., Aspray, W., Yost, J., Tinn, H., & Con Diaz, G. (2023). *Computer. A history of the information machine.* Taylor and Francis. https://doi.org/10.4324/9781003263272

Candès, E., Duchi, J., & Sabatti, C. (2019). *Statistics and the oncoming AI revolution.* https://doi.org/10.1162/99608f92.ff7ca64e

Cerqueti, R., Coppier, R., Girardi, A., & Ventura, M. (2022). The sooner the better: Lives saved by the lockdown during the Covid-19 outbreak. The case of Italy. *The Econometrics Journal, 25*(1), 46–70. https://doi.org/10.1093/ectj/utab027

Chodacki, J., Hanhel, M., Iacus, S., Scherle, R., Olson, E., Pfeiffer, N., Holmes, K., & Hosseini, M. (2024). *GREI Data Repository AI Taxonomy.* https://doi.org/10.48550/arXiv.2411.08054

Cipollone, P., & Rosolia, A. (2007). Social interactions in high school: Lessons from an earthquake. *The American Economic Review, 97*, 948–965. https://doi.org/10.1257/aer.97.3.948

Desai, T., Ritchie, F., & Welpton, R. (2016). *Five safes: Designing data access for research* (Technical Report 1601. University of the West of England, Bristol Economics Working Paper Series). https://doi.org/10.13140/RG.2.1.3661.1604

Diel, A., Lalgi, T., Schröter, I. C., MacDormon, K., Teufel, M., & Bäuerle, A. (2022). Human performance in detecting deepfakes: A systematic review and met-analysis of 56 papers. *Computers in Human Behavior Reports, 16*(100538), 1–13. https://doi.org/10.1016/j.chbr.2024.100538

Dwork, C. (2008). Differential privacy: A survey of results. In M. Agrawal, D. Du, Z. Duan, & A. Li (Eds.), *Theory and applications of models of computation* (pp. 1–19). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-79228-4_1

EC. (2020). *Commission Recommendation on a common Union toolbox for the use of technology and data to combat and exit from the COVID-19 crisis, in particular concerning mobile applications and the use of anonymised mobility data.* (EU) 2020/518 of 8 April 2020. Archive: http://data.europa.eu/eli/reco/2020/518/oj.link

EC. (2023). *Commission communication on long-term competitiveness of the EU: Looking beyond 2030.* COM(2023)0168 of 16 Mar 2023. Archive: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52023DC0168

ECJ. (2018). *Judgment of the Court (Second Chamber) of 20 December 2017*. Peter Nowak v Data Protection Commissioner. Archive: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:62016CJ0434

EDPB. (2019). *Opinion 3/2019 concerning the Questions and Answers on the interplay between the Clinical Trials Regulation (CTR) and the General Data Protection regulation (GDPR).* Retrieved 1 Feb, 2026, from https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-art-70/opinion-32019-concerning-questions-and-answers_en

Farrell, E., et al. (2023). *European Data Spaces: Scientific insights into data sharing and utilisation at scale* (Technical report, Publications Office of the European Union, Luxembourg). https://doi.org/10.2760/400188

Gambacorta, L., Jappelli, T., & Oliviero, T. (2025). *Exploring household adoption and usage of generative AI: New evidence from Italy Technical Report 1298* (BIS Working Papers). Retrieved 1 Feb, 2026, from https://www.bis.org/publ/work1298.htm

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Guidolin, M., & Manfredi, P. (2023). Innovation diffision processes: Concepts, models and predictions. *Annual Review of Statistics and its Predictions, 10*, 451–473. https://doi.org/10.1146/annurev-statistics-040220-091526

Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H. A., Epstein, D. H., Leggio, L., & Curtis, B. (2021). Bots and misinformation spread on social media: Implications for COVID-19. *Journal of Medical Internet Research, 23*(5), Article e26933. https://doi.org/10.2196/26933

Hradec, J., Craglia, M., Di Leo, M., De Nigris, S., Ostlaender, N., & Nicholson, N. (2022). *Multipurpose synthetic population for policy applications* (Technical report, Publications Office of the European Union, Luxembourg). https://doi.org/10.2760/50072

Jacobzone, S., Picalarga, S., Dynes, L., Williams, E., Greenwell, H., & Bowers, P. (2025). Unleashing the policy potential of rigorous impact evaluation and randomised trials (OECD Working Papers on Public Governance). *OECD Working Papers on Public Governance, 87* 1–57. https://doi.org/10.1787/5c135873-en

Jones, C. I. (2022). The past and future of economic growth: A semi-endogenous perspective. *Annual Review of Economics, 14*, 125–152. https://doi.org/10.1146/annurev-economics-080521-012458

Kurz, C. F. (2022). Augmented inverse probability weighting and the double robustness property. *Medical Decision Making, 42*(2), 156–167. https://doi.org/10.1177/0272989X211027181

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon Books.

OECD. (2024). *Recommendation of the council on artificial intelligence* (OECD/LEGAL/0449). Retrieved 1 Feb, 2026, from https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

OECD, Bcg, & INSEAD. (2025). *The adoption of Artificial Intelligence in firms: New evidence for policymaking* (Technical report). OECD Publishing. https://doi.org/10.1787/f9ef33c3-en

Paruolo, P. (2005). Automated inference and the future of Econometrics: A comment. *Econometric Theory, 21*(1), 78–84. https://doi.org/10.1017/S0266466605050061

Pinotti, P. (2017). Clicking on heaven's door: The effect of immigrant legalization on crime. *The American Economic Review, 107*, 138–168. https://doi.org/10.1257/aer.20150355

Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications, 11*(1), 1–14. https://doi.org/10.1057/s41599-024-03811-x

Valgarðsson, V., Jennings, W., Stoker, G., Bunting, H., McKay, D. D. L., & Klassen, A. (2025). A crisis of political trust? Global trends in institutional trust from 1958 to 2019. *The British Journal of Political Science, 55*(e15), 1–23. https://doi.org/10.1017/S0007123424000498

van der Laan, M., & Rose, S. (2011). Targeted learning: Causal inference for observational and experimental data. In *Breakthroughs in statistics* (1 ed.). Springer.

Fabio Berton is Policy Analyst on impact evaluation at the Competence Centre for Microeconomic Evaluation at the Joint Research Centre of the European Commission. He is also Associate Professor of Economic Policy at the University of Torino, and Research Fellow at CIRET, IZA and LABORatorio R. Revelli. His research focuses on applied labour economics, and on the evaluation of labour policies in particular. Fabio has publications on The Economic Journal, The British Journal of Industrial Relations, The Oxford Bulletin of Economics and Statistics, The Review of Financial Studies and The Journal of Economic Inequality among the others.

Paolo Paruolo is an econometrician, currently Head of the European Commission Competence Centre on Microeconomic Evaluation since 2016. He holds a Ph.D. in Econometrics from the University of Copenhagen, and he has been Full Professor of Econometrics at the University of Insubria, Varese, Italy, and previously at the University of Bologna. His research focuses on econometric theory, time series analysis and counterfactual methods. He has published in Econometrica, Journal of Econometrics, Econometrics Journal, Econometric Theory, Journal of the Royal Statistical Society Series A, Oxford Bulletin of Economics and Statistics, among others. He has written and edited several books, including "Data-driven policy impact evaluation: How access to microdata is transforming policy design", Crato, Paruolo (eds 2019).

Ian Vollbracht is a Deputy Head of Unit at the Joint Research Centre (JRC) of the European Commission in Ispra, Italy. He has worked for the EU institutions for the past 25 years, including posts in DG ECFIN, DG TRADE and the Cabinet of Commissioner Cecilia Malmstrom. An economist by training, his current research interests include the impact of artificial intelligence and big data on economic structures, as well as counterfactual evaluation methods and their application to EU policy. He also serves on the JRC's Editorial Review Board.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

