Nuno Crato
Paolo Paruolo
Editors

Data-Driven Policy Impact Evaluation

How Access to Microdata is Transforming Policy Design

EXTRAS ONLINE

Springer Open

Data-Driven Policy Impact Evaluation

Nuno Crato • Paolo Paruolo
Editors

Data-Driven Policy Impact Evaluation

How Access to Microdata is Transforming Policy Design

Springer Open

Editors
Nuno Crato
University of Lisbon
Lisbon, Portugal

Paolo Paruolo
Joint Research Centre
Ispra, Italy

ISBN 978-3-319-78460-1 ISBN 978-3-319-78461-8 (eBook)
https://doi.org/10.1007/978-3-319-78461-8

Library of Congress Control Number: 2018954896

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

Preface

Policymaking is undergoing profound transformations, thanks to the availability of better data and the possibility to uncover causal effects by using improved statistical methods (namely, counterfactual impact evaluation techniques). This book aims to document these changes in a simple language, from a European perspective. The central ideas of the book can be summarised in four paragraphs.

Firstly, statistical methods now exist to rigorously evaluate the impact of policy measures, even when data do not come from controlled experiments. Micro-econometricians and statisticians specialising in these counterfactual impact evaluation methods are well aware of both their potential and limitations. The scope for rigorous causal analysis that they offer is, however, not widely understood. Owing to the inherent complexity of society and the many concurrent factors that influence an outcome, decision-makers often doubt that it is possible to uncover clear causal relationships.

Secondly, to evaluate policy impacts, it is crucial to have data on the basic units targeted by a policy, i.e. the so-called target group. Similar data need to be available for a comparable set of units, called the control group. This often translates into a need to access microdata, i.e. data at the level of individuals, households, businesses or communities.

Thirdly, microdata are available from several sources; one such source that offers numerous advantages is administrative data (or registry data), i.e. data collected by public entities for their service activities. Data from different registries can be linked together, and also linked with external information, such as data from surveys or data collected by private businesses, to obtain comprehensive datasets that are suitable for policy research. If these datasets are properly organised, policy evaluations can be performed in real time.

Fourthly, the use of microdata faces challenges: some real, others imaginary. There are issues with database quality, linkage and preservation of anonymity. Most obstacles can be overcome with appropriate organisation and modern computer science methods. There are also problems of a political nature; to overcome them, one needs to be aware of all the technical solutions that can be used to keep data safe, as well as to be willing to subject policies to reality checks and to learn from those checks.

To properly analyse these topics from all necessary angles, this book collects contributions from a group of researchers, practitioners and statistical officers who work in the field, mostly based in Europe: 11 European countries are represented. The pool of contributors is complemented by experts from the Organisation for Economic Co-operation and Development and from Eurostat, who report cases drawn from their organisations' particular experiences.

The various articles in this book consider different policy areas. They include employment, health, professional training, social security and educational issues. Many of the contributions explain and apply various counterfactual econometric methods, serving as a useful reference for their study and use.

The book provides a panorama of microdata issues relating to policy research, of administrative data availability, of various existing systems that can facilitate safe data use, of successful studies on policy impact and even of policy changes made on the basis of such studies. We hope it will be useful to a large readership in Europe and in the rest of the world.

Policymakers and public policy scholars will find here various examples of successful policy evaluation. They will find arguments in favour of data-based policy studies and a clear case for improving the effectiveness of policy measures.

Public administrators and technical staff at public administrations will find systematic examples showing that policy evaluation is a viable task, with dedicated methods and sophisticated techniques. Various chapters show in detail the reasons why a causal evaluation is possible and the conditions under which these analyses can provide conclusive evidence on the effects of policies.

Statisticians and econometricians will find various discussions on the applicability of counterfactual impact evaluation methods and detailed case studies that show how various techniques can be applied with success. All readers will find practical examples of the most commonly used of these techniques, along with a discussion of their applicability. Statistical officers and database experts will find a state-of-the-art review of anonymisation issues and techniques and of database linkage and security.

To sum up: Microdata exist and can be safely organised for a better knowledge of society and for policy research; data-based policy assessment, monitoring, evaluation and improvement are feasible and desirable, and they can be achieved in unprecedentedly fast ways by properly linking administrative data.

The first chapter introduces these issues. The remaining contributions to this book are grouped into four parts, which are described below.

Microdata for Policy Research

The first part of this book deals with data issues. Paul Jackson provides an updated account of the recent history of the use of microdata. It is only recently, around the beginning of the twenty-first century, that the value of microdata and administrative data started to become fully appreciated. In fact, using microdata from administrative sources for policy research is still innovative. Previously, microdata collected for general statistics were only used to compute summary statistics such as average income, or the number of unemployed people in a particular region. Original records were treated as completely confidential and then discarded.

Now, microdata are understood to have enormous potential for enabling us to improve knowledge about policies, and to a large extent they are inexpensive, as they have been already collected. Official statistical authorities, both at national level and in Eurostat, make microdata available for research. Public administrations such as unemployment agencies and educational registers are increasingly urged by social researchers, political parties and independent organisations to make available the data that they collect regularly and, nowadays, keep in digital format.

Ugo Trivellato explains in what sense microdata should be seen as a public good and makes the case for data availability. He highlights the importance of data release, both for research and for democratic participation. It is a matter of public interest that data should be made available for a better knowledge of society and for the evaluation of public policies. In addition, he reviews some recent advances in regulations and practices on microdata access—at the level of EU and several member states, and at transnational level—and concludes that remote data access is the most effective mode for safely sharing highly informative confidential data.

The full use of administrative data requires the linkage of various databases. Natalie Shlomo explains how modern statistical techniques can be applied to data linkage and how probabilistic linkage can bypass some data faults and still provide perfectly usable datasets. She also shows how probabilistic linkage requires appropriate statistical techniques account for linkage errors when carrying out statistical modelling in linked data.

There are, however, concerns about safeguards around microdata. Some are imaginary and may be used as pretexts to keep data out of reach and unused for policy research and public information. Some are real: confidentiality is a major one. Giovanni Livraga explains how modern computer science techniques are able to anonymise datasets and still provide the relevant information for social research.

Microdata Access

Some countries and institutions are world leaders in the use of administrative data. They have already organised platforms and systems to make microdata available in a systematic and safe way. It is both reassuring and inspiring to learn how this has been achieved.

Eurostat is a general statistical provider and a supplier of microdata for research. Aleksandra Bujnowska describes how Eurostat serves as an entry point for safely accessing microdata provided by national statistical offices in the European Union.

In providing access to microdata, the Institute for Employment Research (IAB) in Germany is one of the most successful institutions in Europe. Dana Müller and Joachim Möller describe in detail how the IAB processes social security data, linking it to survey data and organising safe access for researchers to this wealth of information. The numbers of users and of studies stemming from this access are growing steadily: in 2016, there were more than 1000 users and 500 projects.

In many European countries, access to administrative data is hindered by outdated legislation, by security issues and by confidentiality concerns. It is very interesting to see how Hungary has entered the twenty-first century debating a change in legislation about linkage and access to microdata, and how this has been successfully implemented. Ágota Scharle, who then headed the Finance Ministry efforts to change the laws and the practices, explains how successful negotiations led to a very open and modern system of accessing administrative data.

Other inspiring examples come from the Netherlands. Marcel Das and Marike Knoef discuss an innovative infrastructure, the LISS panel, which helps researchers to conduct field surveys and offers the possibility of integrating the survey results with existing administrative data.

Counterfactual Studies

Counterfactual impact evaluations stemmed originally from labour and educational economics; several contributions to this book come from these fields. In the first of these, Pedro Martins describes the design, implementation and evaluation of a job search support and monitoring programme that has been in place in Portugal since 2012, using a regression discontinuity design on administrative data.

Another evaluation of a labour market intervention is presented by Enrico Rettore and Ugo Trivellato, who analyse the Italian public programme called ‘Liste di Mobilità’, which handles collective redundancies in the Veneto region. They use a crescendo of regression discontinuity design techniques, following an increase in the availability of administrative sources on this programme over the course of 15 years. They emphasise that the administrative data had been there from the start, showing how advances in policy research are linked to increased trust between researchers and public administrators.

Home ownership and debt on house have implications for job mobility. Andrea Morescalchi, Sander van Veldhuizen, Bart Voogt and Benedikt Vogt present an analysis of the impact of negative home equity on job mobility in the Netherlands, using a Dutch administrative panel for the period 2006–2011. They use panel fixed effects and find that negative home equity has a moderate negative effect on the probability of changing jobs.

Numerous programmes are active in any given country at any given time, and their evaluation is not simple. For the Netherlands, Rudy Douven, Laura van Geest, Sander Gerritsen, Egbert Jongen and Arjan Lejour present some of the counterfactual work being performed at the CPB Netherlands Bureau for Economic

Policy Analysis. They provide four examples: the first concerns the labour participation of parents with young children, the second deals with tax shifting by small corporations, the third evaluates teacher quality and student achievement and the fourth analyses performance-based pay in mental health care. In these examples, CPB combines the strengths of structural models and of policy impact evaluation methods, including differences-in-differences, regression discontinuity and random assignment with large datasets. Furthermore, they emphasize the importance of good communication strategies of the results to policy-makers.

For Denmark, Jacob Nielsen Arendt and Mette Verner present a study on the long-term effects of a social intervention for young people with conduct disorder problems. Register data enable the authors to construct a comparison group of young people who are receiving alternative social treatments but who have similar parental characteristics and a similar life-cycle profile in terms of previous social interventions and healthcare use. Using propensity score matching, the authors find that participants are more likely than similar young people to take primary school exams, but they have lower upper-secondary education completion rates and lower employment rates; additionally, they are more dependent on welfare income, and they are more often convicted of crimes.

For Italy, Claudio Deiana and Gianluca Mazzarella investigate the causal effect of retirement decisions on well-being. They exploit the exogenous variation provided by changes in the eligibility criteria for pensions that were enacted in Italy in 1995 and in 1997 to compute an instrumental variable estimate of the causal effect. They find a sizeable and positive impact of retirement decision on satisfaction with leisure time and on frequency of meetings with friends. Their results are based on a mix of survey and administrative data.

A final study on the economics of education in Chile is provided by Julio Cáceres-Delpiano and Eugenio Giolito, who investigate the impact of age of school entry on academic progression for children in Chile, using a regression discontinuity design. Thanks to the use of a very detailed administrative database, they are able to find that a higher age at entry has a positive effect on grade point average and on the likelihood of passing a grade, although this impact tends to wear off over time. Children whose school entry is delayed are also more likely to follow an academic track at secondary level.

Use of Results

Policy impact evaluation creates knowledge of what worked, for whom and when. This information can be organised by policy areas to summarise the state of play in the field. Béatrice d'Hombres and Giulia Santangelo present the state of play on counterfactual evidence for active labour market policies in Europe. Matej Bajgar and Chiara Criscuolo explain how the impacts of a major vocational training programme can be evaluated with the use of linked administrative and survey data, drawing on the example of the Modern Apprenticeships in Scotland.

In a different area, Rita Santos, Sara Barsanti and Chiara Seghieri discuss the use of administrative data in the health sector to evaluate the impact of primary care pay-for-performance schemes in England and in Italy.

The final chapter, by Sven Langedijk, Ian Vollbracht and Paolo Paruolo, looks at the future of microdata access across Europe and lays out the benefits and issues involved in increases in the use of administrative data both within and across European countries.

Final Remarks

The idea of the book was conceived at the Competence Centre for Microeconomic Evaluation (CC-ME) in the European Commission Joint Research Centre (JRC) in Ispra, Varese, Italy. CC-ME is a centre for research that supports the European Commission and European Member States on impact evaluation and promotes the use of causal impact evaluation methods.

Many chapters and aspects of the book stem from the work of CC-ME: several authors are part of it, and most of the other authors have cooperated in one way or another with the Centre's activities. The editors are indebted to all authors for their wholehearted collaboration.

Special thanks also go to other CC-ME researchers and colleagues at the JRC who acted as reviewers and provided suggestions and took part in discussions about the book. In particular, thanks go to Massimiliano Bratti, Maurizio Conti, Claudio Deiana, Leandro Elia, Sophie Guthmuller, Corinna Ghirelli, Massimiliano Ferraresi, Enkelejda Havari, Athanasios Lapatinas, Gianluca Mazzarella, Andrea Morescalchi, Giulia Santangelo, Sylke Schnepf and Stefano Verzillo.

Finally, the project owes a great deal to the unwavering support of Sven Langedijk, the head of the Modelling, Indicators and Impact Evaluation Unit, to which CC-ME belongs. Many thanks go to the European Commission Joint Research Centre for financially supporting open access to the book.

Ispra, Italy
August 2018

Nuno Crato
Paolo Paruolo

Contents

Part I Microdata for Policy Research

Part II Microdata Access

The Power of Microdata: An Introduction

Nuno Crato and Paolo Paruolo

1 Data and Policy

Policy-making is a process guided by ethical values, diverse interests and evidence. It is motivated by political convictions, limited by available resources, guided by assumptions and supported by theoretical considerations. It is also bound by reality checks, which are sometimes reassuring, at other times bring unexpected results, but which are in all cases beneficial.

Economists and social scientists have theoretical models that help assess the intended effect of policies. Given policy goals, these models guide the choice of intervention, such as public investment, changes in the reference interest rate or reformulations of market regulations. However, theory has its limits and can clash with reality.

In modern democracies, a comparison of expressed intentions with actual results is increasingly required by citizens, the media, political groups and policy-makers alike, and rightly so. As Milton Friedman, the 1976 Nobel Laureate in Economics,

This chapter draws freely on work in progress at the European Commission Joint Research Centre (JRC); see Crato (2017). The authors thank Sven Langedijk and researchers at the JRC Competence Centre on Microeconomic Evaluation for useful comments and discussion. The views expressed in this paper do not necessarily reflect those of the European Commission.

N. Crato (✉)
Joint Research Centre, Ispra, VA, Italy

University of Lisbon, Lisboa, Portugal
e-mail: ncrato@iseg.ulisboa.pt

P. Paruolo
Joint Research Centre, Ispra, VA, Italy
e-mail: paolo.paruolo@ec.europa.eu

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_1

once said, ‘One of the great mistakes is to judge policies and programs by their intentions rather than their results’.

Recent years have seen the rise of a ‘what works’ approach to the policy cycle, in which policy interventions are designed using elements that have worked in the past and are evaluated quantitatively to measure their impact.[1] This is happening in parallel with a ‘credibility revolution’ in empirical economics, which Angrist and Pischke (2010) describe as the current ‘rise of a design-based approach that emphasizes the identification of causal effects’.

Public policy can derive benefit from two modern realities: the increasing availability and quality of data and the existence of modern econometric methods that allow for a causal impact evaluation of policies. These two fairly new factors mean that policy-making can and should be increasingly supported by evidence.

The remaining sections of this chapter briefly introduce these two realities: on the one hand, the availability and use of microdata, especially of the administrative type, and, on the other hand, the main modern counterfactual econometric methods available for policy evaluators. A short Glossary completes the chapter.

2 Data Granularity

The granularity of data plays an important role in building evidence for policy. Granularity ranges from ‘micro’, as in microdata, which usually relate to individuals, firms or geographical units, to ‘aggregate’, for state-level data, as in national accounts. Data of different granularities are good for different policy evaluation purposes. Microdata are especially fit for finding evidence of a policy intervention’s effectiveness at the individual level, while aggregate data are useful for studying macroeconomic effects.

As an example, consider a programme of incentives for post-secondary vocational training and its evaluation, during or after its implementation. It is usually assumed that these incentives help to attract young people to technical professions that increase their employability.

One might first think to use the number of youngsters enrolled in such training programmes and examine the aggregate unemployment rate of the cohorts which include people exiting these programmes. This approach would, however, present a number of pitfalls.

Firstly, it would be difficult to know whether a change in enrolment or in youth unemployment was due to general economic conditions or to the programme under analysis. Secondly, one would not be able to directly link employment with the training programme: it might be that the newly employed people were just those

[1] Examples of this approach are the What Works Network in the United Kingdom (https://www.gov.uk/guidance/what-works-network) and the What Works Clearinghouse in the United States (https://ies.ed.gov/ncee/wwc/). All web links in this chapter were last accessed in October 2017. See also Gluckman (2017).

who had not attended the training programme. In summary, aggregate employment rates (even when broken down by cohorts) would not provide evidence of a causal link between the programme and the employment rate.

Suppose now that individual data (microdata) have been collected and that, for each young person eligible for the incentive programme, one knows whether or not he or she has applied to the programme and received the incentives, whether or not he or she has successfully concluded the training provided (treatment) and whether or not he or she has obtained a job (outcome of interest). On top of this, suppose one knows other individual characteristics, such as age, gender, education, parents' occupations and family socio-economic status; these characteristics are examples of 'control variables'.

Finally, assume that all this information is available for both the young people that accessed the incentives, i.e. the treated group, and for people of the same age with similar characteristics who did not follow the programme, i.e. a potential control group. If one could assume that the difference between the treated and the control group was not systematic, as reflected by their age and other individual characteristics (controls), then one could measure directly the success of the incentive programme and assess its impact. (A comparison of the employment rates of the two groups would deliver the average treatment effect of the incentive programme.)

This example shows how microdata, unlike aggregate data, can allow one to identify the impact of a policy. To access such information it is necessary to record it in the first place. Data then need to be linked to follow people throughout the whole relevant period. Next, data need to be made available for the study to be performed. Specific issues are involved at each stage.

3 Administrative Data

Administrative data (admin data) are data collected for administrative purposes by governments or other public administration agencies in the course of their regular activities. Admin data usually consist of large datasets containing, for example, in the case of individuals, data on taxes, social security, education, employment, health, housing, etc. Similar public archives exist containing data on firms or data on municipalities.

These datasets are extensively and continuously updated. They are used for general official purposes, such as control of payments or administrative actions. Recently, they have been recognised as an important data source for policy research and policy impact evaluation, see Card et al (2010).

Given the scope and extent of these databases (some of which may fall into the 'big data' category), there are several advantages for policy research in using admin data, possibly in combination with survey data. Firstly, the quality of the data is in some aspects superior to the one of the data made available via surveys, because the data are maintained and checked for administrative purposes; this results in greater accuracy, which is particularly important.

Secondly, the data usually cover all individuals, firms or municipalities present in the whole population, and hence, the database is much larger than the samples used in surveys.[2] Thirdly, as they coincide with the reference population, they are representative in the statistical sense. Moreover, they do not have or have fewer problems with attrition, non-response and measurement error than traditional survey data sources.[3]

Moreover, admin data have other additional non-negligible practical advantages. Fourthly (adding to the previous list), data have already been collected, and so costs are usually limited to the extraction and preparation of records. Fifthly, data are collected on a regular basis, sometimes on a real-time basis, so they provide sequential information to build time series. Sixthly, data are collected in a consistent way and are subject to accuracy tests. Seventhly, data collection is not intrusive in the way that surveys are. Finally, data linkage across registries is possible and often straightforward, whenever individuals have unique identifiers, such as national ID numbers. Admin data can also be linked to survey data.

Admin data also have limitations with respect to surveys and other types of data collected for specific research purposes. Firstly, the variables recorded may fail to include information relevant for research. Secondly, data reliability may be suboptimal for some variables that are not of central concern for the administrative tasks. Thirdly, data collection rules may vary across periods and institutions. All this implies that admin and survey data may complement each other for a specific purpose.

During the past 15 or 20 years, interest in admin data for social research and policy evaluation has been increasing exponentially—see Poel et al. (2015), Card et al. (2015) and Connelly et al. (2016)—especially when they are complemented by other types of data, including big data; see Einav and Levin (2014) for a general discussion on how large-scale datasets can enable novel research designs.

In a process that began with some occasional uses in North America (see Hotz et al. 1998) and Europe, the wealth of admin data and the possibilities they offer have been increasingly recognised in the past two decades. The call to action in the United States reached the National Science Foundation (Card et al. 2010; White House 2014; US Congress 2016), which established a Commission on Evidence-Based Policymaking, with a composition involving (a) academic researchers, (b) experts on the protection of personally identifiable information and on data minimisation

[2] The resident population of a municipality may be taken to be a (random) sample from a larger fictitious population of similar municipalities; hence, the use of data from the whole resident population does not invalidate the problem of statistical inference.

[3] Attrition refers to the possibility that surveyed individuals may stop participating in the survey. Non-response to the survey refers to people or firms not agreeing to be interviewed. This may imply that respondents self-select in ways that create bias in results; for instance, more successful firms may be more willing to respond than less successful ones. This is called non-response bias and it is a form of selection bias. Finally, measurement error refers to the possibility that the interviewer expects certain answers, which may introduce bias (interviewer bias), that respondents may not recall facts correctly (recall bias), etc.

and (c) policy-makers from the Office of Management and Budget. Its final report, CEP (2017), provides a vivid overview and outlook on evidence-based policymaking in the US.

There have been similar developments in Europe with regard to the use of admin data for policy research purposes, albeit with heterogeneity across states. Some countries already make considerable use of admin data for policy research.[4] The European Commission (2016) issued a directive establishing that data, information and knowledge should be shared as widely as possible within the Commission and promoting cross-cutting cooperation between the Commission and member states for the exchange of data for better policy-making.

In parallel with this progress, researchers have developed methods for improving data quality, data linkage and safety of data access and use. Data quality has been improving continuously in Europe as a result of a set of factors, namely, a continuous effort to make data classification criteria uniform, better monitoring of spending of European Union (EU) funds, increasing attention to regulation efficiency and an intensification of accounting information control over individuals and firms.

Record linkage has also progressed in many countries and has evolved into a highly technical task that has its own methods and issues; see Winkler (2006) and Christen (2012). In the collection of admin data, it makes good sense to establish routines for data linkage. Data are made available to researchers and public institutions in a way that protects confidentiality; there are ways of establishing safeguarding rules, legal standards, protocols, algorithms and computer security standards that make it almost completely certain that relevant data are accessed and studied without violating justifiable confidentiality principles (see, e.g. Gkoulalas-Divanis et al. 2014; Aldeen et al. 2015; Livraga 2015).

For scientific reproducibility (see Munafò et al. 2017), citizens' scrutiny, policy transparency, quality reporting and similar goals, it is also desirable that essential data that support studies and conclusions are made available for replication (reproducibility) or contrasting studies.

A report by President Obama's executive office (White House 2014) considers 'data as a public resource' and ultimately recommends that government data should be 'securely stored, and to the maximum extent possible, open and accessible' (p. 67). The previously cited communication to the European Commission of November 2016 also contains a pledge that, where appropriate, 'information will be made *more easily accessible*' (p. 5).

[4] See, for example, http://fdz.iab.de/en.aspx (Germany), http://www.dst.dk/en/TilSalg/Forskningsservice# (Denmark), https://snd.gu.se/en/data-management/register-based-research (Sweden), https://www.cbs.nl/en-gb/corporate/2017/04/more-flexible-access-to-cbs-microdata-for-researchers (the Netherlands) and the review in the OECD (2014).

4 Counterfactual Methods

Human society is the result of such complex interactions that many people consider it almost impossible to assess the real effect of policies. Indeed, the evaluation of policies' impact is fraught with difficulties; however, it is not impossible.

During recent decades, statisticians and econometricians have been developing techniques that allow for sound conclusions on causality. The better the data used, the sounder the conclusions can be. These methods build and expand on the methods used to analyse experimental data; these extensions are crucial, as microdata are often collected in real-life situations, rather than in experimental environments.

Coming back to the example of training incentives, the evaluation of the policy impact aims to answer the question: 'What would have happened if this intervention had not been put in place?' In natural sciences, this type of question can often be answered by conducting an experiment: in the same country and for the same population, two almost identical groups would be formed, and the policy measures would be put in place for one of the groups (the treated group) and not the other (the control group). The two groups could be formed by random assignment.[5]

Only in rare social situations, however, can a controlled experiment be conducted. There may be objections, for example, on ethical grounds: a deliberate experiment may even be considered discriminatory against one of the groups. Outside controlled experiments, other problems arise. For instance, if individuals or firms self-select into policy interventions, this may change the reference populations for the treated and control groups and cause the so-called selection bias problem.

Notwithstanding all this, a reasonable answer to the same counterfactual question can be achieved with a judicious application of appropriate statistical techniques, referred to as counterfactual impact evaluation (CIE) methods. These methods are called quasi-experimental, because they attempt to recreate a situation similar to a controlled experiment.

CIE methods require data and specific linkages of different databases. Going back to the example previously discussed, the best way to study the effects of the programme would be to follow individuals and record their academic past, their family background, their success in the programme and their employment status. The relevant ministry of education might have data regarding their academic track record, a European Social Fund-funded agency might have data regarding people

[5] The large number of observations (sample size) associated with access to microdata can ease statistical inference. As an example, consider a statistical test of equality of the averages of the treated and controls, which provides a scientific check of the effectiveness of the policy. The power of the test, i.e. the probability to reject the null hypothesis of equal averages when they are different—i.e. when the policy is effective—is an increasing function of the sample size. Therefore, the abundance of microdata improves power of policy research.

enrolled in the training programme, and the relevant social security department might have data regarding (un)employment. These data would need to be linked at the individual level, to follow each individual through the process.

5 Counterfactual Impact Evaluation Methods

In controlled experiments, the average of the outcome variable for the treated group is compared with that for the control group. When the two groups come from the same population, such as when assignment to both groups is random, this difference estimates the average treatment effect.

In many real-world cases, random assignment is not possible, and individuals (or firms) self-select into a treatment according to observable and unobservable characteristics, and/or the selected level of treatment can be correlated with those characteristics. CIE methods aim to address this fundamental selection bias issue.[6]

Some of the standard classes of CIE methods are briefly introduced below in non-technical language. Many excellent books now exist that present CIE methods rigorously, such as the introductory book by Angrist and Pischke (2014) and the books by Imbens and Rubin (2015) and by Angrist and Pischke (2009).

5.1 Differences in Differences

This CIE technique estimates the average treatment effect by comparing the changes in the outcome variable for the treated group with those for the control group, possibly controlling for other observable determinants of the outcome variables. As it compares the changes and not the attained levels of the outcome variable, this technique is intended to eliminate the effect of the differences between the two populations that derive from potentially different starting points.

Take, for example, an impact evaluation of the relative impacts of two different but simultaneous youth job-training programmes in two different cities. One should not look at the net unemployment rate at the end of the programmes, because the starting values for the unemployment rate in the two cities may have been different. A differences in differences (DiD) approach instead compares the magnitudes of the changes in the unemployment rate in the two cities.

A basic assumption of DiD is the common trend assumption, namely, that treated and control groups would show the same trends across time in the absence of policy

[6] This is sometimes referred to as an endogeneity problem or an endogenous regressor problem, using econometric terminology; see, for example, Wooldridge (2010), Chapter 5.

intervention. Hence, the change in the outcome variable for the control group can be used as an estimate of the counterfactual change in the outcome variable for the treated group.

5.2 Regression Discontinuity Design

This CIE technique exploits situations in which eligibility for the programme depends on certain observable characteristics, such as a requirement to be above (or below) an age threshold, such as 40 years of age. Individuals close to the threshold on either side are compared, and the jump of the expected outcome variable at the threshold serves as an estimate of the local average treatment effect.

As an example, consider an EU regulation that applies to firms above a certain size; regression discontinuity design (RDD) can be used to compare the outcome of interest, such as the profit margin, of treated firms above but close to the firm-size threshold with the same figure for control firms below but also close to the firm-size threshold. Firms that lie around the cutoff level are supposed to be close enough to be considered similar except for treatment status.

RDD requires policy participation assignment to be based on some observable control variable with a threshold. RDD is considered a robust and reliable CIE method, with the additional advantage of being easily presentable with the help of graphs. Since the observations that contribute to identifying the causal effect are mainly those around the threshold, RDD may require large sample sizes.

5.3 Instrumental Variables

Instrumental variable (IV) estimation is a well-known econometric technique. It uses an observable variable, called an instrument, which predicts the assignment of units to the policy intervention but which is otherwise unrelated to the outcome of interest.[7] More precisely, an instrument is an exogenous[8] variable that affects the treatment (relevance of the instrument) and the outcome variable only through its influence on the treatment (exclusion restriction).

For instance, assume one wishes to evaluate whether or not the low amount of R&D expenditure in a country is a factor hampering innovation. A way in which this question can be answered is by considering an existing public R&D subsidy

[7] Selection for treatment may depend on unobservable factors that also influence the potential outcomes; this is called selection on unobservables. IV and DiD (for unobservables that are time invariant) can solve the selection bias problem, under rather mild assumptions.

[8] Here, exogenous means an external variable that is not affected by the outcome variable of interest; see Wooldridge (2010) for a more formal definition.

to firms. Assume that in this specific case, subsidies have been assigned through a two-stage procedure. In the first stage, firms had to apply by presenting projects; in the second stage, only those firms whose projects met certain quality criteria were considered (Pool A). Within Pool A, a randomly selected subgroup of firms received the subsidy, as public resources were not sufficient to finance all the projects.

In this scenario, the evaluators can collect data on each firm in Pool A, with information on their amounts of R&D expenditure (policy treatment variable), the number of patent applications or registrations (outcome of interest) and an indicator of whether or not they were given the subsidy. This latter indicator is an instrument to assess the causal effect of R&D spending on innovation (e.g. the number of patent applications or registrations).

Receiving the subsidy presumably has a positive effect on the amount of R&D spending (relevance). Receiving the subsidy is exogenous, since the subsidies were allocated randomly and not according to a firm's innovation potential, which may have caused an endogeneity problem, and is expected to affect innovation only via R&D effort (exclusion restriction).

There is a vast econometric literature on IV, which spans the last 70 years; see, for example, Wooldridge (2010).

5.4 Propensity Score Matching

This CIE technique compares the outcome variable for treated individuals with the outcome variable for matched individuals in a control group. Matching units are selected such that their observed characteristics (controls) are similar to those of treated units. The matching is usually operationalised via a propensity score, which is defined as the probability of being treated given a set of observable variables.[9]

As an example, imagine that one needs to evaluate the impact of an EU-wide certification process for chemical firms on firms' costs. This certification process is voluntary. Because the firms that applied for the certification are more likely to be innovative enterprises, one should compare the results for the treated firms with those for similar untreated firms. One possibility is to define the control group by matching on the level of R&D spending.

Propensity score matching (PSM) requires a (comparatively) large sample providing information on many variables, which are used to perform the matching.

[9] Propensity score matching requires that, conditionally on controls, the potential outcomes are as good as randomly assigned, a condition called conditional independence assumption (CIA) or selection on observables.

6 A Call to Action

As briefly summarised in this chapter, it is now possible to make significant advances in the evaluation and readjustment of public policies. A wealth of admin data are already being collected and can be organised, complemented and made available with simple additional efforts.

Admin data allow for better, faster and less costly studies of economies and societies. Modern scientific methods can be used to analyse this evidence. On these bases, generalised and improved studies of public policies are possible and necessary.

At a time when public policies are increasingly scrutinised, there is an urgent need to know more about the impact of public spending, investment and regulation. Data and methods are available. Data collection and availability need to be planned at the start of policy design. It is also necessary to systematically evaluate the evolving impact of policies and take this evidence into account. In the end, citizens need to know how public investment, regulation and policies are impacting upon their lives.

Appendix: A Short Glossary

- **Administrative data**—Data collected by government entities and agencies in the course of their regular activity for normal administrative purposes, such as to keep track of attendances, tax payments, hospital visits, etc. Administrative data, or **admin data**, are not collected for research purposes. At the moment of collection, these data have high level of granularity, as information is gathered at the individual level.
- **Aggregate or collective data**—Data kept at the general, i.e. summary, level, providing statistics such as totals or averages for the whole population or for sectors of the population.
- **Anonymisation**—A process of guaranteeing that the use of data records for specific and normally temporary purposes does not allow the identification of the individual units in the database.
- **Big data**—A generic expression denoting modern large datasets available in digital format from a great variety of sources. The usual characterisation of **big data** relies on the so-called three Vs: volume, variety and velocity (Laney 2001). The volume of data currently being collected, kept and analysed is unprecedented and makes it necessary to use specific methods for their study; the variety of sources, from administrative records to web use records and from bank transactions to GPS use records, has been made possible only in recent times; and the velocity at which data is gathered approaches real time. These characteristics of much modern data create new opportunities for improving the

lives of citizens, but they also entail serious challenges involving, for example, confidentiality and data treatment.

Arguing that false data have no value, some researchers claim that these three Vs are insufficient and add another: veracity. The resulting four Vs have the backing of the IBM Big Data & Analytics Hub. More recently, other experts in the field (Van Rijmenam 2013) have proposed the seven Vs, adding variability, visualisation and value.

- **Biometrics**—In the field of data analysis, biometrics refers to a process or data that can be used to identify people by one or more of their physical traits.
- **Causality**—The sufficient link from one factor or event, the cause, to another factor or event, the effect. In econometric methods, a plausible establishment of causality requires some type of experiment or the construction or identification of some counterfactual situation (see 'Counterfactual impact evaluation (CIE)') that allows a reasonable comparison of what happened in the presence of a given factor with what happened or can be reasonably accepted as likely to have happened in the absence of the same given factor.
- **Confidentiality**—Restriction of the pool of persons who have access to particular information, usually individually identifiable information. The concept is different from that of respecting private or sensitive information.
- **Control group**—A group adequate for comparison with the group of units that were subject to a given policy (or treatment group, in statistical terminology). Prior to the policy intervention, the control group should display average characteristics that were otherwise similar to those of the group of individuals subject to the measures. The identification of a control group is critical for measuring the effect of a policy intervention, as it indicates what the situation would be for the group subject to the policy intervention had the intervention not been implemented. See also 'Counterfactual impact evaluation (CIE)'.
- **Correlation**—A measure of linear statistical association between variables. The establishment of a reasonable correlation between variables does not imply the establishment of a causal effect, i.e. 'correlation is not causation'.
- **Counterfactual impact evaluation (CIE)**—Refers to statistical procedures for assessing the effect of a policy measure and gauging the degree to which it attained its intended consequences. In randomised control trials, one compares the outcomes of interest of those having benefited from a policy or programme (the 'treated group') with those of a group that are similar in all respects to the treatment group (the comparison or control group) except in that it has not been exposed to that policy or programme. The comparison group seeks to provide information on what would have happened to the members subject to the intervention had they not been exposed to it—the counterfactual case. The difference in the outcome of interest between the treated and control groups provides information about the effect of the policy.
- **Database linkage**—The process of joining information from different databases with information about the same units. For example, an education database may be joined with an employment database to study, at the unit level, the impact of training on employment. Linkage may be performed deterministically

by using unique identifiers (for each unit, information is joined univocally from the different databases) or probabilistically (for each unit, information is plausibly joined, but with admissible and desirably infrequent errors). Linkage may join individual information from various databases, or it may join individual information from one database with contextual aggregated information from other databases.

- **De-identification**—The same as anonymisation.
- **Granularity of data**—The degree of detail of data recorded. The minutest detail is the unit under appreciation. For instance, in a database on tax payments, the highest degree of granularity is attained when data are kept for each person or contributing entity. The term has its origin in atomic physics and computer science.
- **Macrodata**—Usually the same as aggregate data.
- **Metadata**—An explanation of what a given set of data contains, to allow data inventory, discovery, management, evaluation or use. Metadata can be descriptive, if they explain how data can be used and identified; structural, if they explain how data are organised; and administrative, if they describe how the data were created and who can access them.
- **Microdata**—Data collected at the individual level of units considered in the database. For instance, a national unemployment database is likely to contain microdata providing information about each unemployed (or employed) person.
- **Personal data**—Data related to an individual who can be identified from them or from these data and other data that are in the possession of or are available to the data user or data controller. Personal data is a different concept from that of sensitive data.
- **Privacy**—A person's right or privilege to set the conditions for disclosure of personal information.
- **Randomisation**—The assignment of individuals to a group or groups (such as treated and control groups) at random.
- **Reidentification**—The process of combining information from several datasets, by linking them or by using selected partial information, to identify a certain person or entity from previously anonymised datasets.
- **Sensitive data**—Information about an individual, entity, institution or nation that can reasonably be considered harmful if disseminated.
- **Survey data**—Sample data collected for a given purpose from a given population. Usually, survey data are collected from samples constructed with probabilistic methods and so cover only part of the population, although their purpose is to extrapolate the conclusions to the whole universe under consideration. A restrictive definition of the term limits its use to data collected through survey interviews.

References

Aldeen YAS, Salleh M, Razzaque MA (2015) A comprehensive review on privacy preserving data mining. Springerplus 4:694. https://doi.org/10.1186/s40064-015-1481-x

Angrist JD, Pischke J-S (2009) Mostly harmless econometrics: an empiricist's companion. Princeton University Press, Princeton

Angrist JD, Pischke J-S (2010) The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. J Econ Perspect 24(2):3–30

Angrist JD, Pischke J-S (2014) Mastering metrics: the path from cause to effect. Princeton University Press, Princeton

Card D, Chetty R, Martin F, Saez E (2010) Expanding access to administrative data for research in the United States. In: Schultze CL, Newlon DH (eds) Ten years and beyond: economists answer NSF's call for long-term research agendas. American Economic Association, Nashville

Card D, Kluve J, Weber A (2015) What works? A meta analysis of recent active labor market program evaluations. Ruhr Econ Papers 572, RWI Essen, Essen. https://doi.org/10.4419/86788658

CEP (2017) The promise of evidence-based policymaking. Report of the Commission on Evidence-Based Policymaking. https://www.cep.gov/cep-final-report.html

Christen P (2012) Data matching concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer, Heidelberg

Connelly R, Playford CJ, Gayle V, Dibbend C (2016) The role of administrative data in the big data revolution in social science research. Soc Sci Res 59:1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Crato N (2017) A call to action for better data and better policy evaluation. European Commission, Brussels. https://doi.org/10.2760/738045

Einav L, Levin J (2014) The data revolution and economic analysis. Innov Policy Econ 14:1–24

European Commission (2016) Communication to the Commission 'Data, information and knowledge management at the European Commission'. C(2016) 6626 final of 18 October 2016. European Commission, Brussels

Gkoulalas-Divanis A, Loukides G, Sunc J (2014) Publishing data from electronic health records while preserving privacy: a survey of algorithms. J Biomed Inform 50:4–19

Gluckman P (2017) Using evidence to inform social policy: the role of citizen-based analytics. Office of the Prime Minister's Chief Science Advisor, Auckland http://www.pmcsa.org.nz/wp-content/uploads/17-06-19-Citizen-based-analytics.pdf

Hotz VJ, Goerge R, Balzekas J, Margolin F (eds) (1998) Administrative data for policy-relevant research: evaluation of current utility and recommendations for development. Advisory Panel on Research Uses of Administrative Data of the Northwestern University/University of Chicago Joint Center for Poverty Research, Chicago

Imbens GW, Rubin DB (2015) Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge University Press, Cambridge

Laney D (2001) 3D data management: controlling data volume, velocity, and variety. Application Delivery Strategies File 949, META Group Inc., Stamford. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Livraga G (2015) Protecting privacy in data release. Springer, Cham

Munafò M et al (2017) A manifesto for reproducible science. Nat Hum Behav 1:1–9. https://doi.org/10.1038/s41562-016-0021

OECD (2014) OECD expert group for international collaboration on microdata access: final report. Organisation for Economic Co-operation and Development, Paris http://www.oecd.org/std/microdata-access-final-report-OECD-2014.pdf

Poel M, Schroeder R, Treperman J, Rubinstein M, Meyer E, Mahieu B, Scholten C, Svetachova M (2015) Data for policy: a study of big data and other innovative data-driven approaches for evidence-informed policymaking—report about the state-of-the-art. Technopolis Group, Oxford Internet Institute and Centre for European Policy Studies https://ofti.org/wp-content/uploads/2015/05/dataforpolicy.pdf

US Congress (2016) US Public Law 114-140—Mar 30, 2016 'Evidence-Based Policy-making Commission Act of 2016'. https://www.congress.gov/114/plaws/publ140/PLAW-114publ140.pdf

Van Rijmenam A (2013) Why the 3 Vs are not sufficient to describe big data. https://datafloq.com/read/3vs-sufficient-describe-big-data/166

White House (2014) Big data: seizing opportunities, preserving values. Executive Office of the President, White House, Washington https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

Winkler W (2006) Overview of record linkage and current research directions. Bureau of Census Technical Report, Washington https://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

Wooldridge JM (2010) Econometric analysis of cross section and panel data, 2nd edn. MIT Press, Cambridge

Nuno Crato After studying economics at Lisbon Technical University and working as a quantitative consultant for project evaluation and management, Nuno Crato graduated with a PhD in applied mathematics from the University of Delaware and worked in the United States for many years as a college professor and researcher. Now a professor of mathematics and statistics at the University of Lisbon working as a visiting scientist at the JRC, he has published extensively in the fields of time series analysis, econometrics and applied probability models. He served as president of Taguspark, the largest science and technology park in Portugal. An active science writer, he wrote more than a dozen books, some published in the United Kingdom, the United States, Portugal, Brazil and Italy, receiving a prize from the European Mathematical Society in 2003 and a Science Communicator Award from the European Union in 2008. From 2011 to 2015, he was the Portuguese Minister of Education and Science. During his tenure, the dropout rate was reduced from c. 25% to 13.7%, retention rates improved, and Portuguese students achieved the best results ever in international surveys. He has continually pledged for data availability and data-based policy evaluation.

Paolo Paruolo is the coordinator of the Competence Centre on Microeconomic Evaluation (CC-ME) at the European Commission Joint Research Centre, Ispra, IT. He has a master in economics and statistics from the University of Bologna (1987) and a PhD in mathematical statistics (theoretical econometrics) from the University of Copenhagen (1995). He has taught econometrics at the University of Bologna and at the University of Insubria (Varese, IT). His research interests are in econometrics (theory and practice) and in counterfactual methods. Because of his publication record, he was ranked among the best 150 econometricians worldwide in Baltagi (2007) 'Worldwide econometrics rankings: 1989–2005', Econometric Theory 23, p. 952–1012.

Part I
Microdata for Policy Research

From ‘Intruders’ to ‘Partners’: The Evolution of the Relationship Between the Research Community and Sources of Official Administrative Data

Paul Jackson

1 Genesis

If the start of the modern period of research use of official microdata can be dated, then we might decide on 12 June 2003. On that day Julia Lane gave a keynote speech to the Conference of European Statisticians (CES) in Geneva,[1] describing the opportunities and the challenges of using confidential official microdata for research in a new way. Julia encouraged us all to recognise that the complexity of twenty-first century society requires statistical and other data-rich government institutions to work together with the research community in partnership. Julia’s argument was that neither community would be able to meet the challenges on its own, but when working together their different strengths came to more than the sum of their parts. In 2003 a lot of work lay ahead if this partnership was to be possible, let alone successful.

2 Official Data

The focus here is on *official data*, also commonly referred to as ‘administrative data’, meaning the individual records of people and businesses that are obtained by public authorities in order for public services and administration to be carried out. These records are obtained under compulsion or provided in order to use public

[1] http://www.unece.org/fileadmin/DAM/stats/documents/ces/2003/crp.2.e.pdf

P. Jackson (✉)
Administrative Data Research Network, University of Essex, Essex, UK
e-mail: paul.jackson@essex.ac.uk

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_2

services. For Council of Europe member states, this can be seen as an interference with Article 8's right to private and family life—an interference that is justified when necessary in a democratic society and when carried out in accordance with the law.

If the partnership between the research community and the official data community is to be successful, these twin considerations have to be addressed—*lawfulness* and *necessity*. A lot of work has taken place since 2003 on the *lawfulness* of interference with privacy for research use of official data. Perhaps not enough has been done on the matter of *necessity*.

3 Research as a Lawful Interference with Private and Family Lives

The official data that third-party researchers have always turned to first are the data collected and collated by official statistics agencies. In the 1990s, the legal and policy frameworks of official statistics agencies prioritised the integrity of official data to build public and business confidence in published official figures. Guarantees for the confidentiality of official records were a very important part of these reforms. The Conference of European Statisticians drives the statistical work of the United Nations Economic Commission for Europe (UNECE). Its membership includes the European countries but also countries from North America, Australasia and Asia. The CES is a forum for agreeing, explaining and implementing transnational guidelines, standards and reviews of the production of official statistics. This role includes formal global assessments of national statistical systems. Taking their lead from the United Nations Statistical Commission, those global assessments required a benchmark and assisted national systems in achieving it. Thus, the CES had an important role to play in building the integrity of official statistics in its region and spent much of the decade designing and implementing the *Fundamental Principles of Official Statistics*.[2] Member countries had to address its new Principle 6:

> Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

Through the 1990s almost every European country modernised its legal framework for statistical use of official data. Implementation of the Fundamental Principles very often resulted in national legal and policy frameworks that considered research use of official microdata as a threat to Principle 6 and either ruled it out or established very severe controls. This sometimes put researchers into a class of undesirables labelled as 'intruders'. A lot of work went into designing policies and practices to deal with intruders, and sometimes an insufficient distinction was made

[2] http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

between malicious intruders and those in research positions who had no desire to damage the integrity of official data. Nuanced messages are hard to communicate to a sceptical public, and it is much easier for public authorities to give unconditional assurances that nobody else will ever see your private information.

Through this same period, the volume and sensitivity of personal data held by government departments and statistics offices increased, as did the threats to data security. Digital government expanded the amount of information that could be maintained in active service, and to deliver more efficient and more joined-up services, these digital data were increasingly linked, often using a single personal identifier. This improved the power of the information but also put very large volumes of data at risk of unauthorised access and use. The information management and security controls for official data had to be substantially improved, and they were.

The net result of the 1990s' changes in information volume and sensitivity, its management and its legal framework was to increase the gap between the settings applied to the data in government and the settings in place in the research community. Social and economic research in academia, which had most to gain from using the personal digital information held by public authorities, did not progress in a coherent and deliberate way with reforms that mirrored the changes the official data community were undergoing. The official data community was building transparent governance and high-profile public accountability. For example, in the United Kingdom (UK), the government published a consultation document on structural reform to build trust in official statistics in 1998,[3] followed by the *Framework for National Statistics* which introduced the first UK National Statistician role and a code of practice.[4] In the 1990s few, if any, European countries appointed a high-profile and publicly accountable champion of social and economic research with duties equivalent to those of the new National Statisticians, appointments which might have been made under a 'national research law'. Academia did not create and then bind itself to a national code of practice for the research use of official personal information or make a common undertaking to citizens about how confidentiality was to be respected. No regulator of social and economic research and researchers emerged. In the 1990s much of this did happen in the field of health and pharmaceutical research, but not in social and economic research.

When in 2003 Julia Lane asked us all to address these issues, the CES agreed to take on the challenge. Work began to enable the benefits of partnership with researchers without compromising Principle 6. New laws, policies and practices were needed. The CES commissioned a task force to produce principles and guidelines of good practice on managing statistical confidentiality and microdata

[3] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/260823/report.pdf

[4] https://www.statisticsauthority.gov.uk/archive/about-the-authority/uk-statistical-system/history/key-historical-documents/framework-for-national-statistics.pdf

access. They were published in 2007,[5] in what has proved to be a very influential report. The report asserted that when certain conditions are met, the research use of confidential official data is not a breach of Principle 6. The task force's elaboration of these conditions, and the solutions that achieve them, has given us the conceptual building blocks that are elaborated in the legal and policy frameworks in place today across the UNECE region. Since 2003, almost every UNECE country has modified its legal and policy frameworks to implement the task force recommendations and the parallel policy initiatives. New tools and techniques have been developed to take advantage of the new policies and legislation. The main thrust of this work has been to enable conditional gateways through the non-disclosure laws and policies that apply to statistical and other government outputs derived from personal records.

It might be said that research access to official data has followed the same evolutionary change as access to music. At first, if you wanted to hear the music of your choice, you had to play it yourself, and if you wanted data for a particular purpose, you probably had to collect it yourself. Punched cards and then magnetic tape transformed the capture and storage, and reuse, of both music and data at about the same time. Lovers of music started building their own collections of recordings, and researchers started compiling their own copies of data. But then the histories of music and data separated temporarily; for a period the distribution of digital music threatened the property rights of performers and copyright holders, which was something controllers of personal data could never allow to happen. On the research community side, it took a while to accept that having no desire to damage official data's confidentiality is one thing, but inadvertently having that effect is quite another. The local compilation of confidential official data did not become anything more than a great record collection. Thankfully, there was never a 'Pirate Bay'[6] for personal official data. The solution music found was to replace downloads and copying with streaming and digital rights management. Distributing access to content with permissions, rather than distributing the content itself, is inherently safer for all concerned. The equivalent of music streaming in research use of official data is remote access or laboratory access through virtual desktops, and this is now the standard practice.

The UK has this year made primary legislation that provides for the reuse of personal information for research purposes.[7] Appropriately it is part of legislation for a digital economy. It is an example of what is now found in many UNECE countries—the substitution of the original legislative protections under which private personal information was first collected by a public authority with an equivalent but different set of legislative protections that are built around the framework of the 2007 CES Report. Chapter 5 of Part 5 of the Digital Economy Act 2017 provides that whatever statutory or other obligations pertain to the administrative records of

[5] https://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

[6] https://en.wikipedia.org/wiki/The_Pirate_Bay

[7] http://www.legislation.gov.uk/ukpga/2017/30/part/5/chapter/5/enacted

personal information held by a public authority, they are not a legal barrier to the extraction, linking and disclosure of those records to a researcher. The barriers are not simply removed, of course; they are substituted by a set of conditions applying to the researcher, the research project, the parties who prepare the data for the research, the level of anonymisation the data must achieve and the working environment the research must take place in. The UK Statistics Authority is given the statutory duty to establish the criteria for these conditions and the function of accreditation against those criteria.

The UK's Digital Economy Act and its equivalents in other countries are now providing the basis for the lawful interference with private and family lives for the purpose of research. The substitution of one set of protections with another requires the research community to build and maintain the facilities and behaviours necessary to meet the criteria. The Seventh Framework Programme (FP7) project 'Data without Boundaries'[8] explored how the social science data archive community can work in partnership with producers of official data to provide extra capacity and capability for research data access in a form that can achieve accreditation against any reasonable criteria. Data archives, and equivalent social science infrastructures such as the UK Data Service[9] are able to provide accredited research facilities to host official data and to host the approved researchers as they use them. Data archives have capacity and capability advantages that are unlikely to be found in public administration departments. They have the ability to create excellent metadata and data documentation, and they can retain anonymised data for reuse by other projects, saving the multiple extraction costs. Importantly, the providers of many of these services have expertise in linking and anonymisation to create powerful, but very low-risk, research datasets. They have established remote access and laboratory facilities that distribute access, not data. In many countries now, the research community through its funding councils have established world-class services for the preparation and use of research datasets, representing an excellent route for the safe exploitation of data as a key national economic resource. The importance of this service is recognised in the European Strategy Forum on Research Infrastructures (ESFRI)[10] Road Map, with the Consortium of European Social Science Data Archives (CESSDA) recognised as one of its success stories.[11] CESSDA is now a European Research Infrastructure Consortium, with a legal personality and an ability to enter legally binding contracts and receive grants in its own name. CESSDA-ERIC[12] is an example of how the research community is now establishing working environments for the provision of access to data for researchers in accordance with common and accreditable standards in public trust, information security, researcher training, metadata and data documentation and cataloguing.

[8] http://www.legislation.gov.uk/ukpga/2017/30/part/5/chapter/5/enacted

[9] https://www.ukdataservice.ac.uk/

[10] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri

[11] http://www.esfri.eu/esfri_roadmap2016/roadmap-2016.php

[12] https://www.cessda.eu/

CESSDA-ERIC and its members are excellent partners to public authorities who wish to improve lives through the better use of their data, in accordance with the law.

The effect of regulatory legislation can be positive or negative. Those countries subject to the new General Data Protection Regulation[13] may still be undecided as to its effect on research use of official data. Article 6 does not contain a lawfulness provision that unambiguously refers to research in academic and other non-government research institutions, but it does provide for processing in the public interest to be determined lawful. Article 5 provides a positive definition of pseudonymisation which allows for data that have been subject to identity encryption to be used for research purposes without those data being personal data, as long as the encryption key remains beyond the use of the researcher. Article 89 in particular ensures proper recognition of scientific research, suitable technical and organisational measures for protection of any personal data in the information used and an obligation to use data minimisation techniques such as pseudonymisation where possible, to ensure that personal data are used only where necessary. The important concept to retain is that the General Data Protection Regulation is there not to stop the use of personal data for research but to provide a regulatory framework for that processing as an economic imperative in a democratic country pursuing economic well-being.

4 Research as a Necessary Interference with Private and Family Lives

Merely because something is lawful does not make it something that should happen, and the Article 8 right to a private and family life makes this clear. If there is to be interference with privacy, it needs to be both lawful *and* necessary 'in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others'.

Economic research providing evidence that promotes economic well-being can claim a clear provision in Article 8; but only some aspects of social research are similarly provided for. We may regret its absence, but 'the better understanding of society' is not overtly provided for as a just cause for interfering in personal privacy. It would seem sensible to align with the economists and argue the necessity of social research as a factor in the economic well-being of the country.

In turn, it is likely that the public authority most able to make informed national policy to promote economic well-being is also the public authority that holds a great deal of the official data needed to generate the evidence for that policy-making. It follows that the necessity test is met best when a research project is complementary

[13] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

to the analysis of the department. Interference with the privacy of the department's clients may be necessary if the outcome is evidence that fills a knowledge gap in the department or provides elaboration or greater context and insight to the evidence the department already has. The more distant the research product is from the civic functions performed by the public authority, the more difficult it is to argue the interference with privacy is *necessary*. The research community has the capacity and capability to make its work very necessary indeed. Interdisciplinary and longitudinal linking by experts in the research community creates datasets that are potentially more powerful for analysis than the data held by a department acting alone. Researcher training and a career building the skills and knowledge needed to comprehend the messages to be found in the data are to be found across the research communities of all UNECE countries. The challenge to the research community is to ensure that its work is clearly relevant to the economic well-being of the country, and the challenge to the official data community is to recognise that contribution and to value it when it is presented as part of a request for access to personal information.

Both the research community and the official data community want the same thing: better information, for better decisions, for better lives. However, for a long time there have been differences in opinion as to how best to go about it. An official in government lives in fear of data loss, because of the direct responsibility they have to the citizen who gave them their private information and nobody else. A researcher lives in fear of missing a funding window and publication deadlines for prestigious journals. These pressures need not be conflicting, but in practice they too often are. Having the main purpose, the economic well-being of the country in common means that both communities should work hard to bring their perspectives closer together.

From 2011 to 2013, the OECD Expert Group on International Collaboration on Microdata[14] examined the non-legislative barriers to better use of official data. It concluded that a common *language,* a build-up of *trust,* a transparent understanding of *costs versus benefits* and making the provision and use of data for research a *business as usual* activity are every bit as important as the *lawfulness* of data access. The group emphasised that mature partnerships are the best way to address these issues.

The group produced a glossary of terms, in a chapter of its report called 'Speaking the same language'. It is impossible to make a partnership agreement on access to data if the parties to it have different understandings of fundamental terminology. Whole infrastructures can be undermined, even lost, if we are incapable of describing to the public in a coherent manner whether a person can, or cannot, be identified in a research dataset. Initiatives such as the Anonymisation Code of Practice, issued by the UK's Information Commissioner, will only help improve this situation. Our responsibility is to make sure we use these excellent glossaries and guides.

[14] http://www.oecd.org/std/microdata.htm

The OECD group also examined *trust*. Those who wish to use confidential official data have in the past expected the data owner to have faith in them rather than trust in them. Trust is faith with evidence. When a public authority uses their own staff to produce policy evidence, it has reason to believe that this job can be done without risk to privacy or to the integrity of the data. The staff are subject to departmental employment rules; they have been selected at recruitment and trained through their careers; they use official and supported information technology, and they have line management to monitor their conduct. The data are familiar, and any metadata are readily available. The outcomes are known before they are published. The staff know the context of their work and the intention of the policies their evidence is designed to support. Of course, this may also be true of the staff of a research organisation; the question is whether or not there is evidence of equivalence. Unless compelling evidence of equivalence is provided, the data owner cannot have equivalent trust in staff that are not in their employment, trust in information and communications technology and metadata systems that are not under their control or trust in the timing and impact of the publication of evidence. The onus is on the research team to provide that compelling evidence. Independent accreditation may help.

The OECD Expert Group also examined *information as an economic resource*. It is important not to forget the reality of costs and measurable benefits. There is an assumption that the huge budgets of large policy departments and statistics offices can always provide the relatively small resources needed to support extractions of data for research purposes. That may be true, but it is the predictability and notice period of these requests that is important. Budgets, typically, are allocated at the start of the financial year and then spent on the allocated task only. In many departments, in-year flexibility between allocations may be limited. If the part of a department that uses an administrative data source is not allocated a budget for extracting data, building metadata and documentation, attending meetings of the research project team, checking the suitability of the project team and their research environment, checking lawfulness and necessity, etc., then it is not likely to be able to allocate resources for ad hoc project requests. It is incumbent on the research community to co-ordinate its requests for access to data, to compromise on the detail of the data to be extracted from administrative systems and where possible to bundle together a number of projects that can be satisfied with the data that administrative departments are willing and able to provide. This approach offers the most benefit for the least cost, but it does require an infrastructure or another form of strategic leadership to co-ordinate and find compromise on behalf of the community as a whole.

One of the most difficult subjects to discuss is the management of media and public comment on important social and economic research findings. One of the issues addressed in the *Fundamental Principles* and most codes of practice for official statistics is the manner and timing of release of key statistics. Trust in government and its statistics has been low in the recent past. When trust in official statistics is low, it is essential for producers of official statistics to concentrate on their reputation for integrity and trust and therefore to keep the publication agenda on a topic such as crime or poverty, utterly predictable and independent of any other narrative. Pre-announced and independent publication of statistics is essential to prove separation of statistical results from the political narrative, whether the narrative is constructed by government or by single interest groups. The inclusion of published official statistics and other data in research publications does not harm trust in official figures; in fact it probably enhances trust. However, the use of unpublished data from official sources raises a number of public confidence issues. How was it decided which projects would get access to unpublished data and (especially) which would not? Was there interference with the data, or the project, by the provider of the official data? Does the department wish to interfere with the timing and manner of the release of the results for political reasons? On the other hand, does the research project approach a social or economic issue from a particular campaigning position or perspective, selecting for its inquiry an analysis of all the harms—but none of the benefits—of a policy in action? Do the research results tell the markets (or the researcher) anything about the direction of travel of an economic indicator before the official figures are released? It can be difficult to respect the importance of independent research and academic freedom and at the same time maintain public confidence in the timely, predictable, pre-announced and independent release of similar information through official figures. Add a media eager to find information that supports its position on social and economic impacts from political decisions, and it is only a brave and confident national statistician or other administrative data owner who enables the production of research results without knowing how and when those results will be released. It may be beneficial to the two communities to introduce a third. If a partnership includes an ‘evidence intermediary’, being an organisation established to discover, collate and reveal evidence for key areas of policy, then the partnership has its own independent arbiter of what evidence is needed when, and why.

5 A Future in Partnerships

For this latter reason especially and for all the other reasons elaborated here, the research community and the owners of administrative data should seek to enter into partnership agreements. Such agreements would establish the expectations, the contributions, the desired shared outcomes and the details of common actions necessary to get the best evidence from administrative data, thereby enabling the best decisions for the best lives of citizens.

Within partnership agreements we should find, among other things, agreement on:

- The necessity of the research enabled by the partnership to economic well-being
- The lawfulness of the research within the conditions established in legislation
- The accreditation procedures for researchers, projects and their working environment
- The language that will be shared to describe the use of the data

- The costs, and the benefits, the partners expect to experience
- The routine for decision-making through the period of the partnership
- The schedule of work and the production of findings and policy evidence
- The evidence intermediaries that can help identify the most important shared areas of research interest
- How trust will be built and maintained throughout the relationship

The creation of partnership agreements would deliver the spirit of Julia Lane's 2003 keynote address and take us into a new period of genuine collaboration, for the mutual benefit of data owners, data users and the citizens whose lives are affected positively by the intelligent use of data.

Paul Jackson joined the UK Office for National Statistics in 1998. He developed the protocol in the National Statistics Code of Practice for confidentiality and data sharing and established ONS's Microdata Release Panel and Approved Researcher conditions to improve access to ONS confidential microdata. Paul has worked on the European Statistical Law including the regulation for research access to Eurostat's microdata and chaired the OECD Expert Group on international collaboration on microdata. Paul was on the steering committee of the FP7 project 'Data Without Boundaries' and was the first managing director of the Consortium of European Social Science Data Archives (CESSDA). Paul is now the strategic data negotiator for the UK's Administrative Data Research Network, tasked with improving the flow of data from government departments to researchers in a new research council-funded infrastructure.

Microdata for Social Sciences and Policy Evaluation as a Public Good

Ugo Trivellato

The balance between the right to privacy and the right to freedom of information is altered when scientific research comes into play, because of its inherent needs and societal function. This paper argues that, for research purposes, microdata should be characterised as a public good. The evolution of the rules and practices in the European Union (EU) for protecting confidentiality, while allowing access to microdata for research purposes is reviewed. Two key directions are identified for further improvement: remote access to confidential data and the enlargement of the notion of 'European statistics' to include microdata produced for evaluating interventions (co)financed by the EU.

1 Setting the Scene

The issue of access to microdata for research purposes is multifaceted. In fact, it is at the crossroads of two concerns: the right to privacy, on the one hand, and the needs of scientific research on the other. The right to privacy is established in the European Convention on Human Rights and Fundamental Freedoms, reiterated and explicitly extended to the 'protection of personal data' in the Charter of Fundamental Rights

U. Trivellato (✉)
University of Padova, Padova, Italy

FBK-IRVAPP, Trento, Italy

CESifo, Munich, Germany

IZA, Bonn, Germany
e-mail: trivell@stat.unipd.it

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_3

of the European Union (EU).[1] However, this is not an absolute right, as it must be balanced against other competing rights: (1) freedom of expression and information, including freedom ‘to receive and impart information’ (Article 11), where freedom to receive information is considered to imply freedom to seek it; and (2) freedom of the arts and sciences, which affirms that ‘scientific research shall be free of constraint’ (Article 13).

What are the data needs of scientific research per se and for its role in improving the well-being of society? As the Royal Society (2012, p. 8) convincingly argues, ‘open inquiry is at the heart of the scientific enterprise. [It] requires effective communication through [. . .] *intelligent openness*: data must be *accessible* and readily located; they must be *intelligible* to those who wish to scrutinise them; data must be *assessable* so that judgments can be made about their reliability [. . .]; and they must be *usable* by others. For data to meet these requirements it must be supported by explanatory metadata (data about data)’ (emphasis added).

Economic and social sciences[2] face these concerns when data include information primarily on an identified or identifiable person and also on another identified or identifiable agent, such as a firm or an administration.[3] How can these tensions be reconciled? This chapter will take the point of view of an EU-based researcher, focusing on some fundamentals of the issue and their policy implications, rather than on legal and technical aspects.

The rest of the chapter is organised as follows. Section 2 discusses the needs of scientific research and its societal role, in relation to processing microdata. Section 3 summarises the legislation on data protection. Section 4 reviews the evolution of the rules and practices for protecting confidentiality while allowing access to appropriate microdata for research purposes. Section 5 discusses the present state of play in the EU as a whole. The concluding section focuses on the way forward.

2 Scientific Research: Intrinsic Needs and Societal Role

This section outlines the role of individual information in scientific research and points to the growing need for microdata for social science and policy evaluation and stresses the importance of replicability in science. These points are discussed in turn

[1] See Council of Europe (1950, Article 8) and European Parliament (2000, Articles 7 and 8), respectively. The Convention was ratified by all member states of the Council of Europe, among which are those of the EU; the Charter became legally binding with the entry into force of the Treaty of Lisbon, 1 December 2009.

[2] Biological and medical sciences frequently involve personalised intervention and thus face additional challenges, which are out of the scope of this chapter.

[3] In most countries, and in the EU, a concern for confidentiality also extends to the information provided by other units: enterprises, administrations, other institutions, associations, etc. (European Statistical System Committee 2011).

and lead to a characterisation of microdata as a public good, i.e. a non-excludable and non-rivalrous good.

First, the distinctive feature of scientific research is the collective use of individual data. This is elucidated in Recommendation No R (97) of the Council of Europe on the protection of personal data collected and processed for statistical purposes (Council of Europe 1997a).[4] It considers statistics as a scientific discipline that, starting with the basic material in the form of individual information about many different persons, elaborates 'statistical results', understood as characterising 'a collective phenomenon'. This interpretation is extended to fundamental scientific research, which 'uses statistics as one of a variety of means of promoting the advance of knowledge. Indeed, scientific knowledge consists in establishing permanent principles, laws of behaviour or patterns of causality [or patterns of a phenomenon] which transcend all the individuals to whom they apply' (Council of Europe 1997b, p. 7). Moreover, the recommendation points to the need in both the public and private sectors for reliable statistics and scientific research (1) for analysing and understanding contemporary society and (2) for evidence-based decisions. Summing up, statistics and scientific research separate the information from the person: personal data are processed with a view to producing consolidated and anonymous results.

Second, scientific research is experiencing an increasing trend in the use of microdata. Various factors operate to bring about this trend. Some of them act from the supply side, such as technological and statistical advances in data processing, which are making databases of individuals, households and firms more and more widely available. On the demand side, two factors are largely contributing to this trend: (1) from an analytical perspective, the increasing attention paid to individuals (broadly agents), their heterogeneity, micro-dynamics and interdependencies; and (2) the focus on distributive features of policies and on specific target groups of agents, such as in welfare policies and active labour market policies.

In this area, there is a strong demand for assessing the causal effects of programmes, i.e. for estimating the effect of policies on outcomes: this is the core aim of counterfactual impact evaluation (CIE).[5] Correct causal inference depends on knowledge of the characteristics of the population members—the treated and the control groups—relevant to the selection process and the availability of adequate data on them.

The third aspect has to do with replicability, which is essential to science: researchers should be able to rework analyses and challenge previous results using the same data (Royal Society 2012, pp. 26–29). Along the same line, and also

[4] A recommendation is a legal instrument of the Council of Europe, as well as forum organisations such as the OECD and UNECE, that is not legally binding but through the long-standing practice of the member countries is considered to have great moral force. This type of instrument is often referred to as soft law.

[5] The literature on CIE is huge. See, recently, Athey and Imbens (2017). See also European Commission (2013b), a guide for officials responsible for the implementation of European Social Fund-funded interventions.

dealing with CIE, Heckman and Smith (1995, p. 93) stress that ‘evaluations build on cumulative knowledge’. Science is an incremental process that relies on open discussion and on competition between alternative explanations. This holds both for fundamental research and for policy research and implies access to microdata—possibly personal data.

Can the peculiar ‘good’ of personal data processed for research purposes therefore be characterised as a public one?

To answer the question, first consider official statistics, compiled and disseminated by official statistical agencies. Official statistics are a non-rivalrous good. Moreover, collective fixed costs have a dominant role in producing them (Malinvaud 1987, pp. 197–198). But it is not a public good per se, as it is excludable: it would be possible to discriminate among users, both through pricing and through selective access. Thus, characterising official statistics as a public good is a normative issue, the result of a choice in a democratic society. Currently this is a common view: official statistics need to be (and in many countries are) a public good.

Among other things, this view is supported by the principle of ‘impartiality’, one of the Fundamental Principles of Official Statistics adopted by the United Nations Statistical Commission for Europe (UNECE) (UNECE 1992),[6] as well as of the principles for European statistics set out in European Parliament (2009). As stated in the latter, ‘ “impartiality” [means] that statistics must be developed, produced and disseminated in a neutral manner, and that all users must be given equal treatment’.

This argument can be extended to microdata for research purposes, especially when microdata come from public sources or funding (Wagner 1999, Trivellato 2000, among others),[7] provided that (1) eligibility of access is restricted to research purposes and in appropriate ways to researchers, and (2) data access does not compromise the level of protection that personal data require. While intelligent openness remains the paradigm (Royal Society 2012, p. 12), the operational solutions required to achieve it safely remain an issue.

3 EU Legislation on Data Protection

The starting point is Directive 95/46/EC ‘on the protection of individuals with regard to the processing of personal data and on the free movement of such data’ (European Parliament 1995; Directive hereafter). Among its features, two are worth considering.

[6] The Fundamental Principles of Official Statistics, initially adopted by UNECE ‘in the region of the Economic Commission for Europe’, were adopted by the United Nations (UN) Statistical Commission in 1994 and endorsed by the UN General Assembly in 2014.

[7] The contrast between official statistics and microdata has lessened substantially over the past decade.

- Like all EU directives, Directive 95/46/EC is addressed to the member states and requires them to achieve a result—data protection—without dictating the exact means for fulfilling it, thus leaving some leeway. It is up to the member states to bring into force the national law(s) and the administrative provision(s) necessary to comply with the Directive.
- With regard to its scope, the Directive deals with data protection at large, covering almost all kinds of personal data and all of their uses. Thus, it is sparing in offering provisions for their processing for statistical or research purposes.

After a long period of preparation and debate,[8] Regulation (EU) 2016/679 'on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)' (European Parliament 2016; GDPR hereafter), will change significantly the landscape of data protection. The GDPR shall apply from 25 May 2018. Note also that, in contrast with directives, the Regulation shall be binding in its entirety and directly applicable in all member states.

The salient innovations of the GDPR fall under three headings. The first comes with its extended jurisdiction: the GDPR applies to all establishments (companies, public bodies, other institutions, associations, etc.) processing personal data of natural persons residing in the Union, regardless of the establishment's location. The second innovation pertains to the stringent obligations and responsibility of the controller and the processor of personal data[9] (Chapter 4). Finally, the GDPR establishes remedies, liability and penalties in the case of personal data breaches (Chapter 8).

The rest of this section reviews some general provisions of the GDPR and their specifications for the processing of personal data for scientific research purposes.[10] First of all, the GDPR offers a neat definition: '"personal data" means any information relating to an identified or identifiable natural person ([called] data subject)', where an identifiable person is one who can be identified directly (e.g. by reference to an univocal name or an identification number) or indirectly (i.e. by reference to data on one or more factors specific to his physical, physiological, genetic, economic, cultural or social identity) (Article 4(1)).

As for the key principles relating to the processing of personal data, the GDPR stipulates that personal data must be: (a) processed fairly, lawfully and transparently; (b) collected for specified, explicit and legitimate purposes, ordinarily with the

[8] Just the conclusive stage, following the proposal by the European Commission (2012), takes four years.

[9] 'Controller' and 'processor' mean the natural or legal person (or other body) which determines the purposes and means of the processing of personal data and which processes personal data on behalf of the controller, respectively.

[10] The GDPR addresses the 'processing for scientific or historical research purposes or statistical purposes', often jointly with 'the processing for archiving purposes in the public interest' Archiving purposes in the public interest and historical research purposes are irrelevant in the present context. Besides, for the sake of brevity I will use the locution 'scientific research purposes' to refer to scientific research purposes and statistical purposes.

informed, freely given and unambiguous consent of the person, and not further processed in a manner that is incompatible with those purposes; (c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed; (d) accurate and, where necessary, kept up to date; (e) kept in a form which permits identification of the data subjects for no longer than is necessary for the purposes for which the data are processed (principle of 'data minimisation'); (f) processed in a manner that ensures appropriate security of the personal data (Article 5(1)). In addition, the GDPR establishes the information to be given to the data subject, where data have not been obtained from him/her (Article 14), and the rights of the data subject with respect to the processing of his/her personal data: chiefly the rights of access, rectification, erasure ('right to be forgotten'), restriction of processing (Articles 15–18).

When personal data are processed for scientific research purposes, the GDPR determines important derogations to the general provisions. The main exemption is in Article 5(1b), which states that 'further processing for scientific research purposes [of data collected for other specified, explicit and legitimate purposes] shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes',[11] where Article 89(1) stipulates that 'processing for scientific research purposes shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation'. Patently, this exemption is of utmost importance as it allows to process data from registries; indeed, Recital (157) stresses their crucial role in order to facilitate scientific research and to provide the basis for the formulation, implementation and evaluation of knowledge-bases policies.

Additional waivers for the processing of personal data for scientific research purposes, still in accordance with Article 89(1), apply to three more cases.

- While the processing of sensitive data[12] is generally prohibited, the prohibition does not apply when 'processing is necessary [. . . ,] based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject' (Article 9(2j)).
- In the case where personal data have not been obtained from the data subject, the relevant information to be provided to him/her might be substantially reduced, if 'the provision of such information proves impossible or would involve a

[11] The wording is elaborate, with a double denial. Common sense suggests that 'not incompatible' is a synonym of compatible.

[12] The GDPR calls them 'special categories of data'. They comprise 'personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation'.

disproportionate effort, or in so far as the obligation [to provide that information] is likely to render impossible or seriously impair the achievement of the objectives of that processing' (Article 14(4).

- The right to be forgotten shall not apply in so far as the right is likely to render impossible or seriously impair the achievement of the objectives of the processing (Article 17(3)).

Finally, Article 89(2) offers further opportunities for wavers, as 'Union or Member State law may provide for derogations from the rights [of the data subject] referred to in Articles 15 [access], 16 [rectification], 18 [restriction of processing] and 21 [to object,] subject to the conditions and safeguards referred to in paragraph 1 in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes'.

Clearly, substantial room is left to the member states when transposing the Directive into national legislation and to EU institutions for European legislation.

Member states differ appreciably with respect to infrastructure for data collection and dissemination (e.g. national statistical institutes (NSIs) and other statistical authorities and/or social science data archives (DAs), data sources—statistical surveys and/or administrative records). Besides, countries differ with respect to the focus and intensity of the concerns for confidentiality and the ways of handling them, as they are rooted in each country's culture, legislation and practices (Trivellato 2000, pp. 676–681).

Similar observations apply, to a considerable extent, to the GDPR. At the one hand member states will usually incorporate elements of the GDPR in their national law, as far as necessary for coherence and for making the national provisions comprehensible to the persons to whom they apply (Recital (9)). On the other, the rights of a natural or legal person to lodge a complaint with the supervising authority and to an effective judicial remedy against a legally binding decision of a supervisory authority, or against a controller or processor, shall be brought before the supervising authority and the courts of the relevant member state, respectively (Articles 77–79).

4 A Cursory Review of Data Access for Research Purposes in the EU

At the EU level, the process was quite laborious and took a long time, over two rounds: from 1997 to 2002 and from 2009 to 2013. In each round, two regulations were adopted.

Council Regulation No 322/97 (Council of the EU 1997) established the initial framework for the production and dissemination of European statistics,[13] as well as for microdata access for research purposes. On the latter, it states:

1. 'To determine whether a statistical unit is identifiable, account shall be taken of all the means that might *reasonably* be used by a third party to identify the statistical unit'.[14] This does not imply a zero risk of identification; rather, the risk is considered to be practically non-existent when identification would require overly complicated, lengthy or costly operations. Obviously, when statistical units are not identifiable, the microdata set is considered anonymised.
2. Access to confidential data[15] transmitted by the national authorities to Eurostat may be granted by Eurostat itself, provided that it is for scientific purposes and under two further conditions: (b1) explicit approval from the national authority which supplied the data and (b2) enactment of appropriate safeguards for the physical and logical protection of the data.

To draft the subsequent regulation specifically on access to confidential data for scientific purposes, Eurostat was active in promoting an informed debate, with the involvement of NSIs and of members of the research community in advisory committees and working parties and at conferences and seminars (e.g. Jenkins 1999; Wagner 1999; Trivellato 2000; CEIES 2003).[16] Their contributions converged on a guiding principle: capitalising on technological developments and taking appropriate regulatory, organisational and administrative measures (including sanctions), microdata—possibly confidential microdata—should be made available to researchers in accordance with a principle of proportionality (i.e. they should be adequate and not excessive in relation to the purpose) and in a variety of formats. Formats range from 'safe data', i.e. anonymised microdata distributed as public use files (PUFs) to confidential microdata just net of the identifier made accessible to researchers via a 'virtual safe setting',[17] i.e. via safe, remote online access to a secure data storage and computing laboratory within Eurostat and/or a European data archive facility, under appropriate undertakings. In short, this guideline points

[13] In accordance with the Maastricht Treaty in force up to 2009, the regulation refers to 'the community authority' and to 'community statistics'. In this chapter the current wording is used: 'the Commission' or 'the Commission (Eurostat)' or 'Eurostat', where appropriate, and 'European statistics'.

[14] Moreover, the Regulation points out that data taken from sources which are available to the public are not considered confidential (Article 13; emphasis added).

[15] The term 'confidential data' was synonymous with 'personal data' until the promulgation of Regulation (EC) 831/2002, where it takes a quite restrictive meaning. Its meaning is further modified in Regulation No 223/2009.

[16] The European Advisory Committee on Statistical Information in the Economic and Social Spheres, better known under its French acronym CEIES, was set up in 1991 to assist the Council and the Commission in the coordination of the objectives of the EU's statistical information policy. It was replaced by the European Statistical Advisory Committee in 2008.

[17] This mode of access is also known as 'remote data access' or 'microdata online access'.

to the implementation of an adequate set of safe open environments for analysing microdata for scientific purposes, at no (or marginal) cost, and with no appreciable risk of infringing confidentiality.[18]

CEIES (2002) gave significant support to this process: '1. Much significant research in the social and economic spheres, both fundamental and of relevance to the formulation and evaluation of public policies, can only be undertaken with microdata; it cannot be done using published statistics or aggregate records. [. . .] 9. CEIES recommends that Eurostat should establish the feasibility of a virtual safe setting as an alternative to a physical safe setting. If the virtual setting can be put into place, it will be much more cost effective and provide a preferred means of access for the research community'.

Eventually the Commission adopted Regulation (EC) No 831/2002 (Commission of the European Communities 2002), but it took a more conservative stance. Its stated aim was 'to establish, for the purpose of enabling statistical conclusions to be drawn for scientific purposes, the conditions under which access to confidential data transmitted to Eurostat may be granted'. It modified two crucial definitions of the 'father' Council Regulation No 322/97.

1. It established that anonymised microdata shall mean 'individual statistical records which have been modified in order to *minimise*, in accordance with current best practice, the risk of identification of the statistical units' (Article 2, emphasis added). This is at odds with the Council Regulation's criterion based on 'all the means that might *reasonably* be used by a third party'.
2. Previously microdata were considered confidential when they allowed statistical units to be identified, either directly or indirectly, while in this regulation '"confidential data" shall mean data which allow only indirect identification of the statistical units concerned'—which is sensible—and '"access to confidential data" shall mean either access [to proper confidential data] on the premises of Eurostat *or release of anonymised microdata*' distributed under license (Article 2, emphasis added), which is an inconsistent, restrictive adhockery.

The step back with respect to Council Regulation No 322/97 is apparent. Data access was restricted within the conservative paradigm that a balance has to be struck between two conflicting aims, privacy and data needs for scientific research.

The design of the two procedures envisaged in (b) had clear drawbacks. 'Safe data' had to pay the price of a substantial reduction in the information content of the datasets, with the addition burden of obtaining a license. The 'safe centre' on the premises of Eurostat paid the price of severe restrictions placed on access

[18] Based on the experience at the UK Data Archive, for anonymised microdata, Jenkins (1999, pp. 78–81) advocates 'universal access [. . .] for all *bona fide* non-commercial users, subject to registration and standard undertakings of non-abuse, at no cost'. Moreover, he points out that additional components are essential for a sound use of microdata for research purposes: extensive documentation and metadata; information, assistance and training; significant involvement and feedback from analytical users (e.g. via user groups and scientific boards of advisors).

opportunities for researchers, because of the substantial direct and indirect costs incurred by them.[19]

Nonetheless, the availability of microdata from some surveys, granted under Regulation (EC) No 831/2002 via access to the safe centre, turned out to be a significant opportunity. It opened up research to cross-country and (almost) Europe-wide comparisons on significant topics, it helped to create a two-way trust between Eurostat and the community of analytical users, and it contributed to stimulating a growing demand for microdata in the economic and social domains and pushing forward a demand for integration of data from different sources and along the time dimension (e.g. employer-employee-linked longitudinal data).

Moreover, various initiatives by Eurostat, the Organisation for Economic Co-operation and Development (OECD) and UNECE have offered new insights on data access. While advances in computer processing capacity, record linkage and statistical marching open new opportunities for indirect identification, similar developments are also taking place for secure online data access, statistical disclosure control, database protection, disclosure vetting procedures, etc. Overall advances in information and communications technology (ICT) can be harnessed to provide a secure, monitored and controlled environment for access to microdata (e.g. UNECE 2007).

Meanwhile, new potential was emerging from the use of administrative records. Registries draw on the entire (administratively defined) population, are regularly updated and come at no direct cost, which allows the enhancement of statistical and research results (Eurostat 1997, European Commission 2003; see also the series of European Statistical System (ESS) ESSnet projects at https://ec.europa.eu/eurostat/cros/page/essnet_en).

The focus on research data from public funding was another stimulating perspective. In January 2004, the ministers of science and technology of the OECD member countries adopted a *Declaration of access to research data from public funding* and invited the OECD to develop ‘a set of guidelines based on commonly agreed principles to facilitate cost-effective access to digital research data from public funding’. The principles and guidelines, drafted by a group of experts after an extensive consultation process, were endorsed by the OECD Council and attached to an OECD recommendation (OECD 2007).

Lastly, a notable impetus to revision of the EU legislation and practices came from advances made in some member states, such as the Netherlands, the Nordic countries and the United Kingdom (UK). In addition to the diversified practices of safe data dissemination and the establishment of safe centres, between 2000 and 2010, there was a move from piloting to implementation of remote data access services. They include:

[19] This feature demonstrably jeopardised the equality of opportunity of access for researchers. Substantially higher costs were incurred by researchers who travelled a considerable distance to reach the safe setting. Moreover, many academics who needed to combine research with teaching and other obligations were unable to stay at the safe centre long enough to conduct their research.

1. ‘Remote execution’. Registered researchers submit their command files to the safe centre via email, written in one of the admissible statistical packages. At the centre they are moved across a firewall to the server holding the data, and the tasks are run. The results, in the form of output from analyses, are then returned to the researcher by email. This is the case at the LIS Data Center in Luxembourg (home of the Luxembourg Income Study (LIS) and the Luxembourg Wealth Study (LWS) databases; see http://www.lisdatacenter.org) and at some other organisations, including IAB-FDZ, the Research Data Center of the German Employment Agency at the Institute for Employment Research. Note that this mode of access may be cost-effective, but it severely limits the level of interaction of the analyst with the data.
2. ‘Decentralised’ access to a safe centre. Under this mode, accredited researchers, in addition to accessing the data at the safe centre, can do so from ‘safe rooms’ in (a moderate number of) offices which are part of the data provider’s network or at selected universities and other research institutions. For instance, this is the case for the initial provision of decentralised data access in Denmark (Andersen 2003).
3. Proper ‘remote data access’.[20] While its basic format is common to the NSIs and DAs that launched it, procedures and practices vary appreciably in several respects: accreditation procedures, domain of the data made accessible, researcher authentication procedures, output checking, output release, etc. Pivotal cases are remote access at Statistics Denmark (Statistics Denmark 2014, pp. 75–79) and the Microdata ON-line Access implemented at Statistics Sweden starting from the end of 2005 (Hjelm 2006).

Within this renewed interest in extending the accessibility of microdata, the European Parliament (2009) adopted a new Regulation on European statistics, No 223/2009. Known as the ‘Statistical Law’, it marks a profound change. First, it takes a broad, systematic approach to European statistics. It encompasses (1) the reformulation of statistical principles; (2) the reshaping of statistical governance, centred around the notion of the ESS and the role of the ESS Committee in providing ‘professional guidance to the ESS for developing, producing and disseminating European statistics’; (3) the production of European statistics, with provision of access to administrative data sources; (4) the dissemination of statistics; and (5), finally, statistical confidentiality.

No less important are the novelties regarding data access. First, dealing with ‘data on individual statistical units [that] may be disseminated in the form of a *public use file*’, the new Regulation confirms the criterion of taking into account ‘all relevant means that might *reasonably* be used by a third party’ (Article 19; emphasis added). The very same notion of a PUF, and its placement under the heading of ‘dissemination of European statistics’, makes it clear that the set of anonymised

[20] As the number of offices with safe rooms increases and safe rooms are extensively installed in universities and other research institutions, the distinction between decentralised access to a safe centre and remote data access patently fades out.

data is complementary to the set of confidential data and that provisions on data protection do not apply to the former.

Second, the boundary of the confidential data that researchers may access for scientific purposes is sensibly and neatly stated: 'Access may be granted by the Commission (Eurostat) to *confidential data which only allow for indirect identification of the statistical units*'.[21] On the other hand, it remains for the Commission to establish 'the modalities, rules and conditions for access at Community level' (Article 23, emphasis added).

Eurostat implemented various actions for the task and launched several ESSnet projects, such as the feasibility study 'Decentralised Access to EU Microdata Sets' and the project 'Decentralised and Remote Access to Confidential Data in the ESS (DARA)', whose aim was to establish a secure channel from a safe centre within an NSI to the safe server at Eurostat, so that researchers could use EU confidential microdata in their own member states.[22]

Two other important initiatives on transnational access to official microdata were the 'Data without Boundaries (DwB)' project, promoted by the Consortium of European Social Science Data Archives (CESSDA) and launched in May 2011 (http://www.dwbproject.org/), and the 'Expert Group for International Collaboration on Microdata Access', formed in 2011 by the OECD Committee for Statistics and Statistical Policy. The composition of the two teams—both with diversified competences but also with some moderate overlapping—and the constant collaboration between DwB and Eurostat favoured cooperation. Significant results are in OECD (2014), Data without Boundaries (2015) and Jackson (2018).

Furthermore, it is worth to remember that the drafting of a new regulation regarding access to confidential data for research purposes proceeded parallel to—and interacted with—the drafting of the GDPR. Finally, Commission Regulation No 557/2013 was adopted (European Commission 2013a). This was a long way towards reach an appropriate legal framework for access to confidential data for research purposes.

5 The State of Affairs in the EU

By combining the provisions of the two extant regulations,[23] microdata files made available to researchers fit into three categories and four modes of access. The categories are:

[21] If the data have been transmitted to Eurostat, the usual approval of the NSI or other national authority which provided them is required. Note that the restriction to grant access only to confidential data also applies to NSIs and other national authorities.

[22] See https://ec.europa.eu/eurostat/cros/content/decentralised-and-remote-access-confidential-data-ess-dara_en and Gürke et al. (2012), Statistics Denmark (2014), Tubaro et al. (2015).

[23] Additional information is taken from European Commission (2016).

1. *Public use files*: sets of anonymised records of individual statistical units. Provisions for confidential data do not apply to these. On the other hand, no indications are given on how to disseminate them.
2. *Scientific use files*: confidential 'data to which methods of statistical disclosure control have been applied to reduce to an appropriate level and in accordance with current best practice the risk of identification of the statistical unit'. Access is granted to researchers from Member States, European Economic Area (EEA)/European Free Trade Association (EFTA) countries and some EU candidate countries. It takes place in two steps: (b1) recognition of the institution as a research entity and (b2) approval of a research project, submitted by researchers linked to the research entity, who also need to sign a confidentiality undertaking. 'Scientific use files' are then transmitted to the research entity.
3. *Secure use files*: confidential 'data to which no further methods of statistical disclosure control have been applied'. Patently these are the most informative and arguably in many cases are of peculiar interest for research purposes. The accreditation procedure does not vary. But 'access to secure-use files may be granted provided that the results of the research are not released without prior checking to ensure that they do not reveal confidential data'.

Moreover, 'access to secure-use files may be provided only within Commission (Eurostat) access facilities or other access facilities accredited by the Commission (Eurostat) to provide access to secure-use files'. Considering that '"access facilities" means the *physical or virtual* environment [. . .] where access to confidential data is provided', this implies that for secure-use files, two modes of access are envisaged: (1) at Eurostat's safe centre (or another accredited access facility) or (2) via remote data access.

Figure 1, adapted from OECD (2014, p. 8), sketches the secure open environments for accessing the microdata, where the four zones designate datasets with various levels of risk of identification and with which different procedures are associated. Zone 0, the white area outside the circle, refers to anonymised datasets (PUFs), which present a negligible risk of reidentification and are made publicly available, subject to registration and possibly standard undertakings. Zone 1 designates the set of scientific-use files, which entail a moderate risk of identification; they are transmitted to recognised research entities, where they can be accessed by the accredited researchers under adequate security safeguards. Zone 2 designates the set of secure-use files: they entail a high risk of identification and can be accessed only at the access facility itself or via remote data access. NSIs and other relevant national authorities provide directly identifiable personal data to Eurostat in Zone 3: access to them is restricted to Eurostat and the access facility, which perform the set of operations needed to make the confidential data available for research purposes.

This description refers to the 'law on the books'. What about its implementation? The essential results and plans are in Bujnowska (2015, 2016) and at the CROS (Collaboration in Research and Methodology for Official Statistics) Portal

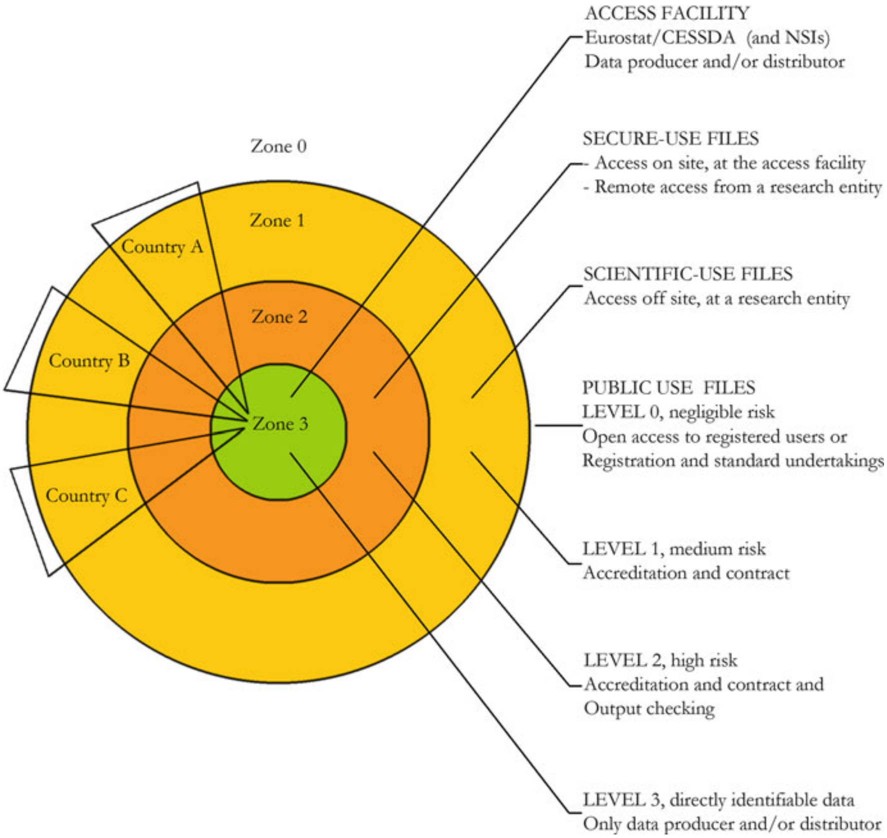

ACCESS FACILITY
Eurostat/CESSDA (and NSIs)
Data producer and/or distributor
Zone 0
SECURE-USE FILES
- Access on site, at the access facility
- Remote access from a research entity
Zone 1
Country A
SCIENTIFIC-USE FILES
Access off site, at a research entity
Zone 2
Country B
PUBLIC USE FILES
LEVEL 0, negligible risk
Open access to registered users or
Registration and standard undertakings
Zone 3
Country C
LEVEL 1, medium risk
Accreditation and contract
LEVEL 2, high risk
Accreditation and contract and
Output checking
LEVEL 3, directly identifiable data
Only data producer and/or distributor

Fig. 1 A set of secure open environments for microdata access for research purposes. Adapted from OECD (2014, p. 8)

Group 'Microdata Access' at https://ec.europa.eu/eurostat/cros/content/microdata-access_en.

Focusing on confidential data,[24] priority has been given to the production and distribution of scientific-use files, also because they demand an extensive, dataset-

[24] As for PUFs, the 'Public use files for Eurostat microdata' project was launched in 2005. In January 2017 Eurostat and CROS released PUFs for the EU Labour Force Survey (2012 and 2013) and the EU Statistics on Income and Living Conditions (2013), for Finland, Germany, Hungary, the Netherlands and Slovenia. Access is open to those registered users of the CROS portal that have also applied for membership of the group on PUFs. Unfortunately, the files are prepared in such a way that individual entities cannot be identified (note that this practice is not in accordance with the identification criterion established by Regulation No 223/2009). This causes a loss in information value and makes the PUFs useful essentially for training and testing only. Hopefully, this restrictive choice is in part—and might be further—compensated by provision of secure access to confidential data.

specific application of statistical disclosure control methods. Results are quite satisfactory. A total of 11 microdata sets had been made available as of December 2016; some 580 research entities have been recognised, and since 2014 more than 300 research proposals per year have been submitted.

As for secure-use files, on-site access has been provided by Eurostat's safe centre in Luxembourg, active for decade, with a comparatively modest investment. Secure-use files are available for the 'Community Innovation Survey' and the 'Structure of Earnings Survey' (2 of the 11 surveys for which scientific-use file versions have been provided) and for the 'Micro-Moments Dataset', an innovative-linked micro-aggregated dataset on ICT usage, innovation and economic performance in enterprises, which enables studies of the economic impact of ICT at company level to be compared across a large sample of European countries.

The use of remote data access secure-use files is attractive for both researchers and Eurostat (or other accredited access facilities), since the microdata do not leave the facility and all output can be controlled. However, no significant advances have so far been made on that front. One crucial reason has been that the envisioned partnership between the ESS and CESSDA had to face a long delay, because of the prerequisite for CESSDA to be recognised as a European Research Infrastructure Consortium (ERIC).

6 Two Suggestions for Improvements

Current initiatives and further steps planned by Eurostat and the ESS for enhancing the use of microdata deal persuasively with various aspects. This section will focus on the need for improvements in two directions: remote data access to secure-use files and reception of a suitably extended meaning of 'European statistics'.

It is no longer controversial that remote data access is an essential ingredient for providing a level playing field for scientific research and for supporting the EU's objective of a 'European research area in which researchers, scientific knowledge and technology circulate freely'.[25] Given the experience of NSIs and DAs in several countries, recently extended to transnational remote data access,[26] it is also largely accepted that remote data access is the most effective mode for sharing highly informative confidential data safely.

The good news is that in June 2017 CESSA became an ERIC. The opportunity for a partnership between Eurostat and CESSDA is now open. This should be a priority for the European Commission and Eurostat. Collaboration should be focused on the facility that will provide the entry point for the EU's microdata access system

[25] Article 179(1) of the Treaty on the Functioning of the EU.

[26] In addition to some pilots carried out within the DARA and DwB projects, it is worth mentioning the Nordic Microdata Access Network, which includes Denmark, Finland, Norway, Sweden, Greenland and Iceland (Statistics Denmark 2014, Thaulow and Nielsen 2015).

(possibly involving NSIs). It should also extend to essential additional components, such as information on ESS microdata products—scientific-use and secure-use files—and any PUFs (e.g. making them discoverable through the resource discovery portal managed by CESSDA); metadata products and services; training and assistance; user conferences and current involvement and feedback from researchers; and production of new microdata files especially for scientific research, which entails the integration of data from different sources or archives and along the time dimension.

The second area where there is a strong demand for improvement falls under the heading 'reception of a suitably extended meaning of European statistics'. First, Regulation No 223/2009 is clear on the need for a more intensive use of administrative records: Eurostat and NSIs 'shall have access to administrative data sources, from within their respective public administrative system, to the extent that these data are necessary for the development, production and dissemination of European statistics' (Article 24). This change may have started, and the production of statistics may also have moved to increased use of administrative sources. But such change is not reflected in increased access to this new data. As pointed out in OECD (2014, Executive summary, Recommendation 51), 'it [is] important to move the information base for microdata access files at the same pace as for statistical production when an office increases its use of administrative data'.

Second, microdata are also produced for monitoring and evaluation of interventions (co)financed by the EU. Their relevance is apparent, as evaluations (at large) are obligatory for all the European Structural and Investment Funds (European Commission 2015) and more emphasis has been placed on CIE, particularly for European Social Fund-funded interventions and research projects. Microdata resulting from interventions as well as from CIE research projects (co)financed by the EU will be made accessible as confidential data for research purposes, preferably as secure-use files via online access to an access facility.

This aim is motivated, and could be implemented, as follows:

1. Microdata produced for monitoring and evaluation should be recognised as part of 'European statistics' and hence included in the European statistical programme. In fact, European statistics are defined as 'relevant statistics necessary for the performance of the activities of the Community' and are 'determined in the European statistical programme' (Regulation No 223/2009, Article 1). Currently microdata produced for monitoring and evaluation are not included in the programme. It is hardly reasonable to deny that they are 'relevant statistics necessary for the performance of the activities of the Community'.[27]
2. Organisations or research units receiving (co)-financing from the EU to carry out evaluations should supply to the European Commission, along with the final

[27] Absurdly, how the Community would be justified to spend money for the production of microdata not relevant and necessary for its activities? From a substantive point of view, one should consider also two further reasons, already referred to, for making these research microdata publicly available: they are data from public funding (OECD 2007); the general argument of Royal Society (2012, p. 8) holds.

report, the full primary data produced, in an appropriate form (i.e. intelligible and assessable, with the relevant metadata). In accordance with the content and the planned use of the microdata, it will be up to the Commission to decide which unit they should deliver this to (e.g. a relevant Directorate-General, Eurostat or the Joint Research Centre).

3. The unit in charge of the management of the microdata should prepare the confidential files—preferably secure-use files—in accordance with the standards set by Eurostat.

It is not a trivial task to specify and implement the above proposal. It would be sensible to consider and discuss these steps promptly.

Acknowledgements The author wishes to thank participants at the seminar held at CRIE-JRC, Ispra, 18 November 2016, for a stimulating discussion and the editors for comments and suggestions on an earlier draft.

References

Andersen O (2003) Access to micro data from Statistics Denmark. In: CEIES (2003), cit., pp 147–152

Athey S, Imbens GW (2017) The state of applied econometrics: causality and policy evaluation. J Econ Perspect 31(2):3–32

Bujnowska A (2015) Access to EU microdata for research purposes. Paper presented at the joint UNECE/Eurostat work session on statistical data confidentiality, Helsinki, Finland, 5–7 October 2015. http://www1.unece.org/stat/platform/display/SDCWS15/Statistical+Data+Confidentiality+Work+Session+Oct+2015+Home

Bujnowska A (2016) ESS microdata access – recent developments. European Commission, Brussels https://circabc.europa.eu/sd/a/6663936c-6704-4fdb-97f1-bccb7608d297/Item%202.1%20Microdata%20access(0).pdf

CEIES (2002) Opinion given on dissemination policy: ‘access to microdata for research purposes’. Eurostat, Luxembourg

CEIES (2003) 19th CEIES seminar: innovative solutions in providing access to microdata, Lisbon, 26–27 September 2002. Publications Office of the European Union, Luxembourg

Commission of the European Communities (2002) Commission Regulation (EC) No 831/2002 of 16 May 2012 implementing Council Regulation (EC) No 322/97 on Community Statistics, concerning access to confidential data for scientific purposes. OJ L 133, 18 July 2002, pp 7–9

Council of Europe (1950) European convention for the protection of human rights and fundamental freedoms. European Court of Human Rights, Council of Europe, Strasbourg http://www.echr.coe.int/Documents/Convention_ENG.pdf

Council of Europe (1997a) Recommendation no R (97) 18 of the Committee of Ministers. Council of Europe, Strasbourg https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680508d7e

Council of Europe (1997b) Explanatory memorandum to recommendation No R (97) 18 of the Committee of Ministers. Council of Europe, Strasbourg https://rm.coe.int/CoERMPublicCommon SearchServices/DisplayDCTMContent?documentId=090000168050a58f

Council of the European Union (1997) Council Regulation (EC) No 322/97 of 17 February 1997 on Community Statistics. OJ L 52, 22 Feb 1997, pp 1–7

Data without Boundaries (2015) EU seventh framework programme, project no 262608. http://www.dwbproject.org/

European Commission (2003) Business register recommendations manual: theme 4 – industry, trade and services. Publications Office of the European Union, Luxembourg

European Commission (2012) Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Brussels, 25.1.2012, COM(2012) 11 final

European Commission (2013a) Commission Regulation (EU) No 557/2013. OJ L 1364, 18 June 2013, pp 16–23

European Commission (2013b) Design and commissioning of counterfactual impact evaluations. Publications Office of the European Union, Luxembourg

European Commission (2015) European structural and investment funds 2014–2020: official texts and commentaries. Publications Office of the European Union, Luxembourg

European Commission (2016) Guidelines for the assessment of research entities, research proposals and access facilities, version 1.4. Publications Office of the European Union, Luxembourg http://ec.europa.eu/eurostat/documents/203647/771732/guidelines-assessment.pdf

European Parliament (1995) Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995. OJ L 281, 23 November 1995, pp 31–50

European Parliament (2000) Charter of fundamental rights of the EU. OJ C 364, 18 December 2000, pp 1–22

European Parliament (2009) Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009. OJ L 87, 31 March 2009, pp 164–173

European Parliament (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. OJ L 119, 4 May 2016, pp 1–88

European Statistical System Committee (2011) European statistics code of practice for the national and Community statistical authorities. Eurostat and European Statistical System, Luxembourg

Eurostat (1997) Proceedings of the seminar on the use of administrative sources for statistical purposes, Luxembourg, 15–16 January 1997. Publications Office of the European Communities, Luxembourg

Gürke C, Schiller D, Gadouche K (2012) Report on the state of the art of current safe centres in Europe. EU Seventh Framework Programme, Project No 262608, Data without Boundaries, Deliverable 4.1. http://www.dwbproject.org/export/sites/default/about/public_deliveraples/d4_1_current_sc_in_europe_report_full.pdf

Heckman JJ, Smith JA (1995) Assessing the case for social experiments. J Econ Perspect 9(2):85–110

Hjelm C-G (2006) MONA – microdata online access at Statistics Sweden. In: Monographs of official statistics: work session on statistical data confidentiality. Publications Office of the European Communities, Luxembourg, pp 21–28

Jackson P (2018) From ‘intruders’ to ‘partners’ – the evolution of the relationship between the research community and sources of official administrative data’. Chapter 2 in this volume

Jenkins SP (1999) Measurement of the income distribution: an academic user’s view. In: Proceedings of the seventh CEIES seminar: income distribution and different sources of income, Cologne, Germany, 10–11 May 1999. Publications Office of the European Communities, Luxembourg, pp 75–84

Malinvaud E (1987) Production statistique et progrès de la connaissance. In: Atti del Convegno sull’informazione statistica e i processi decisionali, Roma, 11–12 Dicembre 1986. Annali di Statistica, Serie IX, vol 7. Roma, Istat, pp 193–216

OECD (2007) OECD principles and guidelines for access to research data from public funding. Organisation for Economic Co-operation and Development, Paris

OECD (2014) OECD expert group for international collaboration on data access: final report. Organisation for Economic Co-operation and Development, Paris

Royal Society (2012) Science as an open enterprise. Royal Society Science Policy Centre, London

Statistics Denmark (2014) Feasibility study regarding research access to Nordic microdata.http://simsam.nu/wp-content/uploads/2016/08/Feasibility-study-regarding-research-access-to-nordic-microdata.pdf

Thaulow J, Nielsen C (2015) New Nordic model for researchers joint access to data from the Nordic statistical institutions. Paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Helsinki, Finland, 5–7 October 2015. http://www1.unece.org/stat/platform/display/SDCWS15/Statistical+Data+Confidentiality+Work+Session+Oct+2015+Home

Trivellato U (2000) Data access versus privacy: an analytical user's perspective. Statistica 60(4):669–689

Tubaro P, Silberman R, Kleiner B et al. (2015) Researcher accreditation: current practice, essential features, and a future standard. EU Seventh Framework Programme, Project No 262608, Data without Boundaries, Deliverable 3.1. http://www.dwbproject.org/export/sites/default/about/public_deliverables/dwb_d3-1_researchers-accreditation_report_final.pdf

UNECE (1992) The fundamental principles of official statistics in the region of the Economic Commission for Europe. http://www.unece.org/fileadmin/DAM/stats/documents/e/1992/32.e.pdf

UNECE (2007) Managing statistical confidentiality and microdata access: principles and guidelines of good practice. United Nations, New York and Geneva

Wagner GG (1999) An economist's viewpoint of prospects and some theoretical considerations for a better cooperation: a German experience. In: Research and development: academic and official statistics cooperation. Publications Office of the European Communities, Luxembourg, pp 89–104

Ugo Trivellato is an Emeritus Professor of Economic Statistics at the University of Padova where he was Professor from 1980 to 2010, FBK-IRVAPP Senior Research Fellow, IZA Fellow and CESifo Fellow. Main research interests are programme evaluation, measurement and modelling of labour supply and unemployment. Previous researches are on structural equation models with measurement errors, data revisions and dynamic economic modelling. Consultant and member of advisory committees of governmental and international agencies on statistical information and microdata access. Publications in various journals include *European Economic Review*, *Journal of Business & Economic Statistics*, *Journal of Econometrics, Journal of the Royal Statistical Society, Labor, Politica Economica, Quality & Quantity, Rivista Internazionale di Scienze Economiche e Commerciali, Statistica* and *Survey Methodology*.

Overview of Data Linkage Methods for Policy Design and Evaluation

Natalie Shlomo

1 Introduction

Data and technology are the building blocks of evidence-based policy. With the increasing availability of government service administrative data and the launch of dedicated research centres such as the United Kingdom (UK) Administrative Data Research Network, the potential for using data and technology to inform policy issues has never been greater.

A key technological tool to exploit the wealth of information contained in administrative data and other data sources is data linkage. In its simplest form, data linkage brings together information from two different records that are believed to belong to the same entity based on a set of identifiers or quasi-identifiers, known as matching variables. As mentioned, the distinction made here is that the aim is to link records for the same entity. This distinguishes the approach taken here from other recently developed methods that have been used to integrate data sources, known as data fusion or statistical matching (D'Orazio et al. 2006).

This chapter focuses briefly on deterministic (exact) matching, where all variables have to match exactly to determine a match. It then focuses on probabilistic data linkage, where allowances are made for errors in the matching variables. There are three possible scenarios in this type of linkage:

- If two records agree on all matching variables, it is unlikely that they would have agreed by chance the level of assurance that the link is correct will be high, and it is assumed that the record pair belongs to the same entity.
- If all of the matching variables disagree, the pair will not be linked as a match, and it is unlikely that the record pair belongs to the same entity.

N. Shlomo (✉)
Social Statistics Department, University of Manchester, Manchester, UK
e-mail: natalie.shlomo@manchester.ac.uk

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_4

- If there are intermediate situations where some matching variables agree and some matching variables disagree, it is necessary to predict whether the pair is a true match or a non-match. Often clerical intervention will be needed to determine the match status.

The challenge in data linkage is when there are errors in matching variables and no unique high-quality identifier such as an ID number is available. In that case, use is made of probabilistic data linkage with matching variables such as name, year of birth or place of residence, which may be prone to error. When combined and concatenated, matching variables should identify an entity uniquely across the data sources (unless the aim is to deduplicate one file). They also need to be accurate and stable over time, so place of residence can be problematic if the timeliness of the two files to be matched is not considered. In addition, there may be differences in how the data is captured and maintained in different databases.

Therefore, the key technical challenges when carrying out a probabilistic data linkage application are the following:

- The availability of good-quality identifiers to discriminate between the entity to whom the record refers and all other entities
- Deciding whether or not discrepancies in identifiers are due to mistakes in reporting for a single entity
- Processing a large volume of data within a reasonable amount of computer processing time

Sections 2 and 3 focus on deterministic (exact) matching and probabilistic record linkage, explained through the three stages of linkage: pre-linkage, linkage and post-linkage. Section 4 describes some recent advances in research related to data linkage, and Sect. 5 provides an overview of methods for the analysis of linked data that may be subject to linkage errors.

2 Deterministic (Exact) Matching Method

In deterministic (exact) matching, the records in two datasets must agree exactly on every character of every matching variable to conclude that they correspond to the same entity. It is generally used when a high-quality identifier such as an ID number is available. If there is no ID number, matching variables, such as age, gender, place of residence, etc. can be concatenated to form a unique ID number, but these variables may be prone to error.

Deterministic matching assumes that there is no error in the way the data is recorded and captured. Information may be missing or inaccurate, and there may be variations in format or inaccuracies in spelling across different sources. For this reason, relaxations have been proposed allowing some errors to be taken into account. For example, first and last names can be transformed into a phonetic code, or the names can be truncated, e.g. by using the first five letters of a name only.

Decision rules can also be set where, if there is an agreement on most of the matching variables, the pair will be declared a match. In deterministic matching, all matching variables have equal weights associated with them, such that an agreement on gender would have the same contribution to the overall decision on a correct match as an agreement on last name, although the latter should clearly contribute more to the decision.

Another important feature is that deterministic matching carries out a one-to-one match, and there are only two possible outcomes of the decision: match or no match. It is useful to carry out a deterministic matching procedure before moving to the probabilistic data linkage if the aim is to carry out a one-to-one match on a set of matching variables, since this may reduce the overall computational burden. Manual review is still needed following deterministic matching to carry out checks for any linkage errors.

3 Probabilistic Data Linkage

A probabilistic data linkage application typically involves three stages:

- Pre-linkage: editing and data cleaning, parsing fused strings such as first and last name or house number and street name, and standardising matching variables so that they have the same formats and definitions.
- Linkage: bringing pairs together for comparison and determining correct matches, i.e. the pair belongs to the same entity.
- Post-linkage: checking residuals for the unmatched, determining error rates and other quality indicators and carrying out analysis taking into account linkage errors.

The following sections describe the probabilistic data linkage method in terms of these three stages: pre-linkage, linkage and post-linkage.

3.1 Pre-linkage Stage

3.1.1 Data Standardisation

The success of data linkage depends on the quality of the data. Pre-processing and data cleaning are the most difficult and time-consuming steps in data linkage, but they are necessary to ensure a successful and accurate linkage. In the first step, a reference number needs to be generated and added to each record across the files to be linked. The reference number should contain a header denoting the iteration of the linkage. Duplicates need to be removed (unless the aim of the linkage is to deduplicate a dataset).

Matching variables need to be selected. The choice depends on the type and contents of the datasets. When no stable ID number is available, it is necessary to link on less stable variables, which may be subject to errors and omissions. The criteria for matching variables are uniqueness, availability, accuracy and stability over time.

Some considerations for matching variables are:

- Proper names rarely change during the lifetime of a person, for example, birth surname, first forename and initials.
- Personal characteristics that are fixed at birth very rarely change, for example, gender, ethnicity, date of birth, place of birth and social security number.
- Social demographic variables may change over time, for example, street name, postcode, marital status, social class and date of marriage.

Matching variables should have high discriminating power. Examples of variables with high discriminating power are those with a large number of value states, such as zip code and last name. Examples of variables with low discriminating power are those with a small number of value states, such as gender and month of birth.

All datasets should be checked for completeness with a clear understanding of the coverage of each of the datasets. Variables involved in the data linkage should be free of errors. Some errors can be detected by checking logical consistencies in the data; for example, a marital status of ‘married’ is not possible for an individual under the age of 14. Ranges of numerical variables and check digits of ID numbers can be easily verified. Other edits and corrections are:

- Matching variables might include fused strings, such as first name given together with last name, or house number given together with street name. These matching variables need to be parsed into separate matching variables, as this increases the power of the decision rule to determine correct matches.
- Names may suffer from variations in spelling, or the use of nicknames and abbreviations, and this increases the complexity of the linkage. This is typically solved by using dictionaries that equate different versions of names and can be tailored to different cultures that use different nicknames, for example, William and Bill. In addition, spelling variations of commonly occurring names and addresses can be replaced with standard spellings using dictionaries. Other errors that need to be addressed are English transliterations of foreign names; the use of initials, truncations and abbreviations; and swapping of surnames and forenames.
- Strings might contain extra words such as ‘Mr’, ‘Mrs’, ‘Dr’, ‘Jr’ or directional ‘East’ or ‘West’. These redundant words are typically removed by a direct linkage to a dictionary containing a list of such redundant words.
- All missing and miscoded data need to have the same definition and notation across the datasets. Missing values have to be consistently labelled as such.
- All files have to have standardised formats for each of the variables to be used for matching, for example, coding of dates should be the consistent across datasets.

- All matching variables must have the same characteristics, field length and coding status across datasets.

3.1.2 Phonetic Codes and String Comparators

To compensate for errors in strings, probabilistic data linkage can make use of phonetic codes and string comparators. Phonetic codes cope with spelling errors, for example 'Reid' and 'Reed', would contain the same phonetic code, as they sound the same and hence would be considered an agreement if they were compared on their phonetic code. The most commonly used phonetic code is Soundex because it exists in many statistical packages. However, for foreign names the code is less satisfactory, since it ignores vowel sounds. There are variations of Soundex that have been developed in different countries. The use of Soundex may, however, cause pairs to agree in a string when in fact they are very different. More robust phonetic codes have been developed. One such code is the New York State Identification and Intelligence System (NYSIIS) code. This code retains information about the position of vowels by converting most vowels to the letter 'A', and it replaces consonants with other, phonetically similar, letters.

String comparator metrics are another way to deal with typographical errors by accounting for deletions, insertions and transpositions (where a letter is moved one position to the left or right) in strings. A string comparator $\Phi_{(S_1,S_2)}$ is a metric between 0 and 1 where 1 denotes a perfect agreement and 0 denotes a perfect disagreement. Jaro (1989) introduced a string comparator that has been shown to be robust when used for data linkage of individuals and is commonly used when linking first and last names. The algorithm is based on the lengths of the two strings denoted by *str_length1* and *str_length2*, the number of common characters across the two strings (the common letter must be within half of the length of the smaller string), denoted *#common*, and the number of transpositions, denoted *#transpositions*. It is defined as follows:

$$\Phi_{(S_1,S_2)} = \frac{1}{3}\left[\frac{\#common}{str\_length1} + \frac{\#common}{str\_length2} + \left(1 - \frac{1}{2}\left(\frac{\#transpositions}{\#common}\right)\right)\right] \tag{1}$$

Winkler (1990) found that fewer errors are made at the beginning of the string than at the end of the string and hence has enhanced the Jaro string comparator by introducing weights. This is known as the Jaro-Winkler string comparator.

Another commonly used string comparator is bigrams. These are typically used in privacy-preserving data linkage where strings are anonymised via computer science functions (see Sect. 4.2). A bigram is two consecutive characters in a string. For example, bigrams in the word 'bigram' are 'bi', 'ig', 'gr', 'ra' and 'am'. The string comparator is defined as the proportion of two character sub-strings in common between the two strings, where the denominator is the average number of sub-strings.

3.1.3 Blocking Variables

In addition to selecting matching variables, there is a need to determine blocking variables. A blocking variable aims to reduce the search space between two datasets by avoiding the comparison of record pairs that are least likely to be matches. For example, if both datasets have 10,000 individuals and a one-to-one match is to be carried out, this results in 100 million pairs to compare. The search space can be dramatically reduced by forming pairs for comparison only among those with the potential to be matches, such as those with a common geographical area. For example, the first three digits of a postcode can be used as a blocking variable. Record pairs are brought together only if they agree (exactly) on the blocking variable. This use of a deterministic matching approach to assist in the probabilistic data linkage greatly reduces the computational burden. However, blocking variables must be as error-free as possible or potential matches can be missed.

The typical approach, especially for one-to-one matching, is to carry out the probabilistic data linkage sequentially, starting with a restrictive deterministic matching on the blocking variable and forming all record pairs for comparison. The pairs are compared, and matches are determined following the method described below. Once matches are determined, they are set aside and a second linkage is carried out through the residual datasets, where the blocking variable can now be less restricted. This is carried out multiple times until the residual datasets are small enough that no blocking variable is needed.

The blocking variables must be small enough to avoid too many unproductive comparisons but large enough to prevent records for the same entity spilling over into adjacent blocks and so failing to compare possible true matches. Therefore, different blocking variables should be used for the iterative passes through the datasets. For example, one might block on postcode and surname, carry out the data linkage and place matches aside and then match residual datasets using a different blocking criteria, such as year of birth or initial of first name and so on.

Once the blocking variable and matching variables are determined for the current iteration, both datasets need to be blocked and sorted and all possible pairs generated. This can easily be done through database Structured Query Language (SQL) commands used for relational database management systems.

3.2 Linkage Stage

3.2.1 Parameters of Data Linkage

In probabilistic data linkage, a frequency analysis of data values is carried out to calculate a weight or score for each matching variable, which indicates how likely it is that they refer to the same entity. Uncommon value agreements should give stronger evidence of linkage. Large weights assigned to matching variables are expected when there is a correct match and small weights assigned to matching variables when there is no match. Note that there is still a positive, albeit small, weight even for an incorrect match, due to the potential for errors in the matching variable. The weight is a ratio of two frequencies:

- Number of agreements of the value of the matching variable in record pairs that represent that same entity (true match)
- Number of agreements of the value of the matching variable in record pairs that do not represent the same entity

This original framework for data linkage was developed by Newcombe et al. (1959) and formalised into the well-known probabilistic framework described in Fellegi and Sunter (1969), referred to hereafter as F&S.

There are three key parameters for probabilistic data linkage, which are represented by the following concepts:

- The quality of the data
- The chance that the values of a matching variable will randomly agree
- The ultimate number of true matches that exist in the database

The quality of the data is represented by the numerator of the above ratio and is denoted as the *m*-probability in the F&S framework: the probability that a matching variable agrees given that the pair is a true match. This is the degree to which the information contained for a matching variable is accurate and stable across time. Data entry errors, missing data or false dates diminish accuracy and produce low-quality data.

The discriminating power of the matching variable is represented by the denominator of the above ratio and is denoted as the *u*-probability in the F&S framework: the probability that a matching variable agrees given that the pair is not a true match. This is similar to the situation where a matching variable will randomly agree across a pair regardless of whether it is a true match or not a true match and is approximately equal to the inverse of the number of values of the matching variable. For example, gender would be expected to randomly agree 50% of the time between pairs and hence does not have high discriminating power.

The third parameter is the ultimate number of true matches or the marginal probability of a correct match. Although this parameter is not explicit in the above ratio, it is essential that there is a sufficiently high proportion of true matches to ensure a successful data linkage.

3.2.2 Basic Concepts in Probabilistic Data Linkage

As mentioned, probabilistic data linkage relies on calculating weights or scores for each matching variable, based on a frequency analysis of the number of agreements as well as disagreements in pairs of records. In the simplest approach, probabilistic data linkage requires some preliminary matching to have been carried out on a similar application. For example, to estimate census undercounts from a post-enumeration survey, the linkage between the survey and the census relies largely on parameters derived from the previous census. Alternatively, a gold standard matched test dataset can be constructed from a small portion of the datasets that have been carefully verified and checked. From this test dataset, the probabilities are calculated that two records will agree on a matching variable in truly matched pairs compared with the probability that records will agree on non-matched pairs or simply by chance. In other words, how likely is it that the variables that agree between a record pair would have done so by chance if the pair was not correctly matched (the u-probability)? This is compared with how likely the agreement would be in correctly matched record pairs (the m-probability). This criterion therefore determines good matching variables, i.e. the agreement between variables should be more typical of correctly matched pairs than of those that might have occurred by chance in unrelated records. Section 3.2.4 describes the expectation-maximisation (EM) algorithm for estimating m- and u-probabilities without the need for test data.

In formal notation in the F&S framework:

For two datasets A and B, let the records in each dataset be denoted $a \in A$, $b \in B$ and the set of all possible matches $A \times B = \{(a, b); a \in A, b \in B\}$. Let $\alpha(a)$ represent the matching variables for entity a in file A and similarly $\beta(b)$ for entity b in file B. The aim is to determine a set of matches $M = \{(\alpha(a), \beta(b)) | \, a = b\}$ and a set of non-matches $NM = \{(\alpha(a), \beta(b)) | \, a \neq b\}$. To develop the decision rule, it is necessary to define a comparison space $C : \alpha(a) \times \beta(b) \rightarrow \Gamma$. This comparison space is composed of a comparison vector $\gamma \in \Gamma$ that represents an agreement pattern (typically 1 for agree and 0 for disagree) for each matching variable. As an example of an agreement pattern for pair j with three matching variables $\gamma^j = \left(\gamma_1^j, \gamma_2^j, \gamma_3^j\right)$, let $\gamma_1^j = 1$ if pair j agrees on last name and 0 otherwise, $\gamma_2^j = 1$ if pair j agrees on first name and 0 otherwise, and $\gamma_3^j = 1$ if pair j agrees on street name and 0 otherwise. One such agreement pattern might be $\gamma^j = (1, 0, 1)$: agree on last name, disagree on first name and agree on street name. In fact, for three matching variables and for a simple agree/disagree {1,0} pattern, the comparison space would contain eight possible agreement patterns. Agreement patterns can also be more complex, using string comparators, for example, $\gamma^j = (0.66, 0, 0.80)$.

The m-probability is now formally defined as the conditional probability that a record pair j has an agreement pattern γ^j given that it is a match (M), denoted as $m = P(\gamma^j | \, M)$, and the u-probability as the conditional probability that a record pair j has an agreement pattern γ^j given that it is not a match (NM), denoted as $u = P(\gamma^j | \, NM)$. Finally, let $P(M)$ be the marginal probability of a correct match.

The probability of interest is the match probability given an agreement pattern γ: $P(M | \, \gamma^j)$. According to Bayes' theorem, this is the posterior probability calculated as follows:

$$\begin{aligned} P\left(M|\gamma^j\right) &= \frac{P\left(\gamma^j|M\right)P(M)}{P\left(\gamma^j\right)} = \frac{P\left(\gamma^j|M\right)P(M)}{P\left(\gamma^j|M\right)P(M)+P\left(\gamma^j|NM\right)(1-P(M))} \\ &= \frac{1}{1+\frac{P\left(\gamma^j|NM\right)(1-P(M))}{P\left(\gamma^j|M\right)P(M)}} \end{aligned} \tag{2}$$

The agreement (likelihood) ratio $R\left(\gamma^j\right) = \frac{P(\gamma^j|M)}{P(\gamma^j|NM)}$ is defined as the test statistic (overall score) for record pair *j*, since maximising the likelihood ratio is the same as maximising the posterior probability of $P(M|\,\gamma^j)$. Therefore, one can simply order the likelihood ratios $R(\gamma^j)$ and choose an upper cutoff W^+ and a lower cutoff W^- for determining the correct matches and correct non-matches. The linkage rule $F : \Gamma \rightarrow \{M, C, NM\}$ maps a record pair *j* comparison value to a set of three classes—matches (M), non-matches (NM) and a set of undecided cases for manual clerical review (C)—defined as follows:

$$F : \begin{cases} \gamma^j \in M \quad if \quad R\left(\gamma^j\right) \geq W^+ \\ \gamma^j \in NM \quad if \quad R\left(\gamma^j\right) \leq W^- \\ \gamma^j \in C \quad otherwise \end{cases} \tag{3}$$

The F&S framework assumes conditional independence across matching variables. This means that the errors associated with one matching variable are independent of the errors associated with another matching variable. Under conditional independence the *m*- and *u*-probabilities can be decomposed as follows: $P\left(\gamma^j|M\right) = P\left(\gamma_1^j|M\right) \times P\left(\gamma_2^j|M\right) \times \cdots \times P\left(\gamma_k^j|M\right)$ and $P\left(\gamma^j|NM\right) = P\left(\gamma_1^j NM\right) \times P\left(\gamma_2^j|NM\right) \times \cdots \times P\left(\gamma_k^j|NM\right)$. The likelihood ratio for record pair *j* becomes:

$$R\left(\gamma^j\right) = \frac{P\left(\gamma^j|M\right)}{P\left(\gamma^j|NM\right)} = \frac{P\left(\gamma_1^j|M\right) \times P\left(\gamma_2^j|M\right) \times \cdots \times P\left(\gamma_k^j|M\right)}{P\left(\gamma_1^j|NM\right) \times P\left(\gamma_2^j|NM\right) \times \cdots \times P\left(\gamma_k^j|NM\right)}$$

Taking the log transformation, the overall score based on the likelihood ratio for record pair *j* is the sum:

$$\begin{aligned} \log\left[R\left(\gamma^j\right)\right] = \log\left(\frac{P\left(\gamma_1^j|M\right)}{P\left(\gamma_1^j|NM\right)}\right) + \log\left(\frac{P\left(\gamma_2^j|M\right)}{P\left(\gamma_2^j|NM\right)}\right) \\ + \cdots + \log\left(\frac{P\left(\gamma_k^j|M\right)}{P\left(\gamma_k^j|NM\right)}\right) \end{aligned} \tag{4}$$

Note that any log can be taken for the transformation and that here the natural log is used.

For example, assume from a previous linkage that the following *m*- and *u*-probabilities were obtained:

P(agree on characteristic x|M) = 0.9 if x = first name, last name, year of birth and 0.8 if x = house number, street name, gender

P(agree on characteristic x|NM) = 0.05 if x = first name, last name, year of birth and 0.1 if x = house number, street name, gender

Assume the following record pair j is to be examined:

Name	Address	Age	Gender
Barbara Jones	439 Elm St	1968	M
Barbara Jones	435 Elm St	1969	F

The agreement vector is γ^j = (agree first name, agree last name, disagree house number, agree street name, disagree year of birth, disagree gender) = (1, 1, 0, 1, 0). When there is a disagreement in a matching variable k, the complement of the likelihood ratio or the disagreement ratio is calculated as $\frac{\left[1-P\left(\gamma_k^j|M\right)\right]}{\left[1-P\left(\gamma_k^j|NM\right)\right]}$. The overall score for record pair j with the agreement vector (1, 1, 0, 1, 0), and based on the likelihood ratio of each matching variable, is:

$$\begin{aligned}\log\left(R\left(\gamma^j\right)\right) &= \log(0.9/0.05) + \log(0.9/0.05) + \log((1-0.8)/(1-0.1)) \\ &\quad + \log(0.8/0.1) + \log((1-0.9)/(1-0.05)) \\ &\quad + \log((1-0.8)/(1-0.1)) = 1.129\end{aligned}$$

Similarly, the overall scores are calculated for all record pairs.

While the m-probability represents the quality of the matching variable and is not dependent on the actual value of the matching variable, this is not the case for the u-probability. The u-probability represents the discriminating power of the matching variable, and hence rare values should provide more weight to the overall score than common values. In addition, since the number of non-matches is very large compared with the number of matches when comparing all possible pairs within blocks, the u-probability $P(\gamma|NM)$ is often approximated by the marginal probability $P(\gamma)$. For example, the u-probability of month of birth is often taken as 1/12 and gender as 1/2. Therefore, in many small-scale applications, the u-probability is calculated as the proportion of value states of the matching variable in a large dataset or across all possible pairs. However, the calculation of the m-probability needs good test data or an approach such as the EM algorithm described in Sect. 3.2.4, since this is calculated as the rate of error among known matches.

The likelihood ratio can be modified to take into account a string comparator. A common approach when using the simple agree/disagree {1,0} comparison vector of the F&S framework is by interpolation. Assume matching variable k is first or last name. The likelihood ratio is modified as follows:

$$R\left(\gamma_k^j\right) = \Phi^j_{(S_1,S_2)}\frac{P\left(\gamma_k^j|M\right)}{P\left(\gamma_k^j|NM\right)} + \left(1-\Phi^j_{(S_1,S_2)}\right)\frac{1-P\left(\gamma_k^j|M\right)}{1-P\left(\gamma_k^j|NM\right)} \tag{5}$$

One can see that if there is a perfect agreement and $\Phi^{j}_{(S_1,S_2)} = 1$, then we obtain the original agreement likelihood ratio, and when $\Phi^{j}_{(S_1,S_2)} = 0$ we obtain the disagreement likelihood ratio. Intermediary values are obtained for the likelihood ratio under partial agreements. Finally, on the new likelihood ratio, the log transformation is taken and added to the likelihood ratios of the other matching variables.

For missing values, one might consider taking a comparator $\Phi_{(S_1,S_2)}$ of $1/k$ where k is the number of categories. For example, a missing value in gender could have a string comparator of $\Phi_{(S_1,S_2)} = 1/2$.

For a quantitative variable such as year of birth, one might want to provide more of an agreement if the difference in year of birth is 1 or 2 years compared with a difference in year of birth of more than 3 years. A possible string comparator is:

$$\Phi_{(S_1,S_2)} = \begin{cases} \exp\left(-|birth\ \ year1 - birth\ \ year2|/3\right)\ if & |\,birth\ \ year1 - birth\ \ year2\,| < 3 \\ 0 & otherwise \end{cases}$$

which obtains a value of 1 if *birth year 1* = *birth year 2*, a value of 0.717 if there is a difference in 1 year, 0.513 if there is a different in 2 years and 0 otherwise.

3.2.3 Setting Thresholds

Based on the observed values of the comparison vector between a record pair, F&S consider the (log) ratio of probabilities in Eq. (4). A decision rule is given by thresholds W^{+} and W^{-} as shown in Eq. (3) which are determined by a priori error bounds. The errors are defined as Type 1 (false matches) and Type 2 (false non-matches) errors. These error bounds are preset by the data linker and should be very small. Generally, the Type 2 error bound is larger than the Type 1 error bound because data linkage is typically an iterative process, with multiple passes through the datasets, and hence missed matched pairs may be found in subsequent passes through the record pairs using different blocking variables. On the other hand, in a Type 1 error, paired records may be obtained that are erroneously declared matches, which can cause severe bias in statistical analysis on the matched dataset. The undetermined record pairs (those record pairs between the upper and lower cutoff thresholds) are sent for clerical review to determine their match status. Fellegi and Sunter (1969) show that, given the error bounds, the decision rule in Eq. (3) is optimal and minimises the number of pairs that need to be clerically reviewed.

The thresholds can be determined by a test dataset where the true match status is known. The empirical distribution of the overall scores from the matches and non-matches are used separately to determine the cutoff thresholds. Before calculating the empirical distribution, the overall scores (likelihood ratios) for each record pair j should be transformed into a probability according to the following transformation:

$p_j = \frac{\exp(W_0+W_j)}{\exp(W_0+W_j)+1}$ where $W_0 = \log\left(\frac{E}{A\times B-E}\right)$ and $A \times B$ is the total number of pairs, E is the expected number of matches and $W_j = \log\,[R(\gamma^j)]$ the overall score calculated in Eq. (4).

Based on the transformed overall scores p_j, we calculate the cumulative empirical distribution of the matches and find the threshold W^- that corresponds to the predetermined Type 2 error bound in the lower tail. We then calculate the empirical distribution of the non-matches and find the threshold W^+ that corresponds to the predetermined Type 1 error bound in the upper tail. All pairs above W^+ are declared matches, all pairs below W^- are declared non-matches and those in between are sent for clerical review, as shown in the decision rule in Eq. (3). Often, the extent of clerical review is determined by available resources. The costs can also be incorporated into the decision rule in Eq. (3) (Pepe 2003).

3.2.4 Expectation-Maximisation Algorithm for Estimating Parameters

Fellegi and Sunter (1969) considered the decomposition of the probability of agreement for record pair j under the simple agree/disagree {1,0} comparison patterns:

$$P\left(\gamma^j\right) = P\left(\gamma^j|M\right)P(M) + P\left(\gamma^j|NM\right)(1 - P(M)) \tag{6}$$

The left-hand side of Eq. (6) obtains the proportion of the agreement patterns across all possible pairs. For example, for three matching variables, there would be 8 ($= 2^3$) possible agreement patterns and hence 8 equations; although since probabilities must sum to 1, the 8th equation is redundant. These probabilities can be used to solve for the probabilities on the right-hand side of Eq. (6). Assuming a simple agree/disagree {1,0} pattern for each matching variable, the m-probability for a matching variable k in record pair j is distributed according to the Bernoulli distribution $P\left(\gamma_k^j|M\right) = m_k^{\gamma_k^j}(1-m_k)^{1-\gamma_k^j}$ and under the assumption of conditional independence across all matching variables: $P\left(\gamma^j|M\right) = \prod_k m_k^{\gamma_k^j}(1-m_k)^{1-\gamma_k^j}$. Similarly, for the u-probability: $P\left(\gamma^j|NM\right) = \prod_k u_k^{\gamma_k^j}(1-u_k)^{1-\gamma_k^j}$. Therefore, for each matching variable k, there are two unknown probabilities, m_k and u_k, as well as the overall match probability, $P(M)$. With three matching variables, seven unknown parameters are obtained. Fellegi and Sunter (1969) showed that, by using information from the frequencies of the agreement patterns on the left-hand side of Eq. (6), one can estimate these unknown probabilities on the right-hand side of Eq. (6).

Data linkage will typically have more than three matching variables; thus the aim of the EM algorithm is to find the best solution. In the E-step, the indicator value is estimated for the true match status denoted by $g_m^j = 1$ if record pair j represents the same entity (set M) or 0 otherwise and $g_u^j = 1$ if record pair j does not represent the same entity (set NM) or 0 otherwise. Applying Bayes' theorem for the simple case of agree/disagree {1,0} agreement pattern and starting with initialising values for the probability of a match (denoted $\widehat{p}$) and m- and u-probabilities for each matching variable, the estimates of the indicator values for the jth record pair are:

$$\widehat{g}_m^j = \frac{\widehat{p}\prod_k \widehat{m}_k^{\gamma_k^j}(1-\widehat{m}_k)^{1-\gamma_k^j}}{\widehat{p}\prod_k \widehat{m}_k^{\gamma_k^j}(1-\widehat{m}_k)^{1-\gamma_k^j} + (1-\widehat{p})\prod_k \widehat{u}_k^{\gamma_k^j}(1-\widehat{u}_k)^{1-\gamma_k^j}}$$

and

$$\widehat{g}_u^j = \frac{(1-\widehat{p})\prod_k \widehat{u}_k^{\gamma_k^j}(1-\widehat{u}_k)^{1-\gamma_k^j}}{\widehat{p}\prod_k \widehat{m}_k^{\gamma_k^j}(1-\widehat{m}_k)^{1-\gamma_k^j} + (1-\widehat{p})\prod_k \widehat{u}_k^{\gamma_k^j}(1-\widehat{u}_k)^{1-\gamma_k^j}}$$

In the M-step, the values of the three probabilities are updated, m-probability, u-probability and the proportion of matched pairs $\widehat{p}$, as follows: $\widehat{m}_k = \frac{\sum_j g_m^j \gamma_k^j}{\sum_j g_m^j}$, $\widehat{u}_k = \frac{\sum_j g_u^j \gamma_k^j}{\sum_j g_u^j}$ and $\widehat{p} = \frac{\sum_j g_m^j}{R}$, where R is the number of record pairs.

These new estimates can be replaced in the E-step and iterated until convergence, i.e. the difference between the probabilities at iteration $t-1$ and iteration t is below a small threshold. One can also plug in the u-probabilities if they are value-specific and known from a large database and use the EM algorithm to estimate the m-probabilities and the overall match probability$\widehat{p}$.

3.3 Post-linkage Stage and Evaluation Measures

After the data linkage process, it is necessary to carry out checks for errors in the match status of the matched and non-matched dataset. A small random sample is drawn from the set of matches and the set of non-matches and the accuracy of the match status is verified, particularly for those record pairs near the threshold cutoff values. These checks allow accurate estimation of the Type 1 and Type 2 errors.

In terms of classic decision theory, the decision matrix for data linkage is presented in Table 1.

The probability of a Type 1 error is the proportion of falsely linked non-matches out of the total number of non-matches. The specificity is the compliment and represents the correctly non-linked matches out of the total number of non-matches.

Table 1 Decision matrix for data linkage

Decision	No match (null hypothesis)	Match (alternative hypothesis)
No link (do not reject null)	Ok (true negative)	Type 2 error (false negative)
Link (reject null)	Type 1 error (false positive)	Ok (true positive)

The probability of a Type 2 error is the proportion of incorrectly non-linked pairs out of the total number of true matches. The compliment is the power of the test, which is the proportion of correctly paired matches out of the total number of true matches. This is also known as the sensitivity. In the data linkage literature, it is also known as recall. Another measure found in the data linkage literature is precision, which is defined as the number of correctly linked matches out of the total number of linked pairs.

In summary:

- Sensitivity/recall—correctly matched pairs out of all true matches
- Specificity—correctly not linking non-matches out of all true non-matches
- Precision—correctly matched pairs out of all possible decisive links

It is important to disseminate these measures to users of the linked data to enable them to understand the quality of the linkage and be able to compensate for linkage errors in statistical analysis and inference (see Sect. 5).

3.4 Constraints on Matching

The essence of a probabilistic data linkage is iterating passes of the datasets in which blocking variables (must match exactly) and matching variables (used to compute the agreement scores) change roles. Blocking variables reduce the computational burden but increase the false non-match rate. Matching variables increase the computational burden and manage the trade-off between false match and false non-match errors. Multiple passes through the pairs are carried out, interchanging blocking and matching variables. As records are linked, they are removed from the input files, and therefore one can use fewer blocking variables to avoid the chance of false non-matches.

In many cases, it may be necessary to match hierarchical datasets, for example, all individuals within households. The characteristics associated with households, such as street name and street number, may overwhelm the characteristics associated with individuals and diminish the power of the statistical test. This is solved by first matching households based on household variables and then matching individuals within households. This partitions the set of pairs as matches within households, non-matches within households and non-matches outside of households. All non-matches are collapsed into one class for the next iteration through the record pairs.

Another requirement is to ensure that matching variables are not highly correlated with each other, since this diminishes the discriminating power of the variable.

For example, age and year of birth should not be used together as matching variables for data linkage.

4 Recent Advances

Two areas that have shown much development in recent years are clustering algorithms for linkage across multiple databases, specifically for deduplication, and privacy-preserving record linkage. More details are provided below.

4.1 Indexing and Blocking

Traditionally, blocking variables partition records based on an exact match to values of a variable such as postcode or first initial of first name (see Sect. 3.1.3), and the data linkage is carried out iteratively, exchanging blocking and matching variables. More formally, this is known as indexing, which is the procedure to reduce the search space of record pairs that are unlikely to contain true matches. Multipurpose indexing can be carried out by performing multiple blocking passes and using different blocking variables and then taking the union of all the retained pairs. Other approaches include a neighbourhood approach, where records are sorted on each database and a sliding window of fixed size is used to define the blocks, and canopy clustering, where a distance metric (e.g. similarity score) is calculated between the blocking variables and records are inserted into one or more clusters. Then, each cluster becomes a block from which record pairs are produced. Filtering discards any pairs not initially excluded by the blocking variables but which are still unlikely to contain a true match. These techniques and others are described in Christen (2012) and Murray (2016). Steorts et al. (2016) and Sadinle (2017) have also applied innovative methods for indexing and deduplication in multiple databases using clustering techniques. This has led to new representations of the data linkage problem within a Bayesian framework.

4.2 Privacy-preserving Record Linkage

In privacy-preserving record linkage, the matching variables are encrypted using computer science techniques instead of coarsened, perturbed or deleted, as is the usual practice in statistical disclosure control. The encryption of matching variables takes the form of ‘hashing’, where the strings are first split into bigrams and then hash functions are used to encrypt each bigram. Another approach is the use of Bloom filters, where strings are encoded into a data structure defined as a bit string (an array of 0s and 1s) that can be represented as an integer in its binary form and used to test whether or not an element is a member of the set (Schnell et al. 2009). Even if matching variables are encrypted, they can be used in exact matching. Furthermore, similarity scores can be used as string comparators; the most commonly used is the Jaccard score. For the hashed bigrams, the Jaccard score is the ratio of the exact matching bigrams divided by the total number of bigrams. For the bloom filter, the Jaccard score is the ratio of the common bits in the string divided by the total number of bits.

As a relatively new area of research, privacy-preserving record linkage is making the crossover from computer science into the statistics community. There is still much work to be done to prove that the method is viable and 'fit for purpose'. Smith and Shlomo (2014) propose data linkage within and across data archives with data custodians allowing the encryption of matching variables based on a common seed. Users can then request a probabilistic data linkage based on the F&S approach through an interface.

One of the drawbacks of privacy-preserving record linkage is that clerical review cannot be carried out except by a trusted third party, who would have access to the original strings and the keys for the encryption.

5 Analysis of Linked Data

Research into the analysis of linked data that are subject to linkage errors has recently gained traction. Earlier work by Scheuren and Winkler (1993, 1997) followed by Lahiri and Larsen (2005) and Chambers (2009) has dealt with the problem of linkage errors in regression modelling. Assume a one-to-one match on two datasets of the same size where a variable X on file A is linked to a variable Y from file B. C is defined as the permutation matrix of {0,1} representing the data linkage. For example:

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_3 & y_1 \\ x_1 & y_2 \\ x_2 & y_3 \end{pmatrix}$$

Consider the regression equation $Y = CX\beta + \varepsilon \quad \varepsilon \sim N(0, \sigma^2 I_n)$. A naive and biased estimate of β is $\widehat{\beta}_N = \left(X'C'CX\right)^{-1} X'C'Y$. If both X and Y are written in their design matrix form, the naive contingency table can be written as $X'C'Y$.

We assume that C is a random permutation matrix whose distribution is dependent on the parametric estimates of the record linkage denoted by ψ and on X. Define $Q = E(C|\,X, \psi)$ as a probability error matrix as follows:

$$\begin{cases} Y_i^* = Y_i \quad with \quad probability \quad q_{ii} \\ Y_i^* = Y_j \quad with \quad probability \quad q_{ij}\; i \neq j \end{cases}$$

where Y_i^* is the linked record and Y_i is the true record.

Lahiri and Larsen (2005) propose an unbiased estimator for the regression parameter β: $\widehat{\beta}_{LL} = \left(X'Q'QX\right)^{-1}X'Q'Y^{*}$. For the contingency table, an unbiased estimator is $X'\ Q^{-1}Y*$. An additional estimator proposed for the contingency table is $X'\ Q'\ Y*$, and although this is a biased estimator, it will generally have smaller mean square error than the unbiased estimator. This is the subject for further research.

How to obtain the error matrix Q is still a subject of open debate. Lahiri and Larsen (2005) suggest using the matching probabilities, themselves derived from the linkage process to define the Q matrix. Chambers (2009) mentions that researchers analysing the data would probably not have access to all the matching probabilities and proposes working with the data linkers to estimate linkage errors using small random samples in a post-linkage step.

Chambers (2009) defines the exchangeable linkage error model for estimating the Q matrix under the following assumptions: one-to-one match and no missed links; data linkage errors occur only within distinct blocks $m,\ m = 1,2\ldots M$, and each block is of size n_m; within each block m, the linkage is non-informative, i.e. the probability of a correct match is the same for all records in the block; and it is equally likely that any two records could in fact be the correct match.

Based on these criteria, the Q matrix is defined as follows:

Let $\Pr(correct\ \ link) = P(C_{m(i,i)} = 1) = q_m$ and $\Pr(not\ \ correct\ \ link) = P(C_{m(i,j)} = 1) = (1 - q_m)/(n_m - 1)$,

where Q is a block diagonal matrix of m blocks of size n_m and in each block q_m is on the diagonal and $(1 - q_m)/(n_m - 1)$ is on the off-diagonal. Note that the row sums to 1, as is the requirement for a probability error matrix. Chambers (2009) also proposed a modification that adapts for false non-matches, which would not appear in the dataset.

In another approach, Goldstein et al. (2012) treat the equivocal links (those links that are not an exact match) as missing values. They then impute the values under two approaches: standard multiple imputation and extended multiple imputation, where the posterior distribution is adjusted by priors defined by the matching probabilities. Under the non-informative linkage assumption (similar to the exchangeable linkage error model) and assuming no model misspecification for the multiple imputation of the equivocal links, the authors achieve good results for reducing bias caused by linkage errors.

Further research is ongoing into other types of regression modelling, the relaxation of the restrictions of the exchangeable linkage error model and also the use of calibration to compensate for missed links. One significant problem that remains is specifying the error probabilities in the Q matrix. It may be possible under certain conditions to simulate the matching process based on the matching parameters and estimate error rates through a bootstrap procedure, as described in Winglee et al. (2005) and Chipperfield and Chambers (2015).

References

Chambers R (2009) Regression analysis of probability-linked data, Official Statistics Research Series 4. Statistics New Zealand, Wellington http://www3.stats.govt.nz/statisphere/Official_Statistics_Research_Series/Regression_Analysis_of_Probability-Linked_Data.pdf. Accessed 2018

Chipperfield JO, Chambers R (2015) Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. J Off Stat 31(3):397–414

Christen P (2012) Data matching: concepts and techniques for record linkage, entity resolution and duplicate detection. Springer-Verlag, Berlin

D'Orazio M, Di Zio M, Scanu M (2006) Statistical matching. Wiley, Chichester

Fellegi IP, Sunter AB (1969) A theory for record linkage. J Am Stat Assoc 64:1183–1210

Goldstein H, Harron K, Wade A (2012) The analysis of record-linked data using multiple imputation with data value priors. Stat Med 31(28):3481–3493

Jaro MA (1989) Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. J Am Stat Assoc 84(406):414–420

Lahiri P, Larsen M (2005) Regression analysis with linked data. J Am Stat Assoc 100:222–230

Murray JS (2016) Probabilistic record linkage and deduplication after indexing, blocking, and filtering. J Priv Confid 7(1):3–24

Newcombe HB, Kennedy JM, Axford SJ et al (1959) Automatic linkage of vital records. Science 130(3381):954–959

Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press, New York

Sadinle M (2017) Bayesian estimation of bipartite matchings for record linkage. J Am Stat Assoc 112(518):600–612

Scheuren F, Winkler WE (1993) Regression analysis of data files that are computer matched – part I. Surv Methodol 19:39–58

Scheuren F, Winkler WE (1997) Regression analysis of data files that are computer matched – part II. Surv Methodol 23:157–165

Schnell R, Bachteler T, Reiher J (2009) Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak 9:41

Smith D, Shlomo N (2014) Record linkage approaches for dynamic database integration: privacy preserving probabilistic record linkage, deliverable 11.1 for WP11, Data Without Boundaries. http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2014-01-Data_without_Boundaries_Report.pdf. Accessed 2018

Steorts RC, Hall R, Fienberg S (2016) A Bayesian approach to graphical record linkage and deduplication. J Am Stat Assoc 111(516):1660–1672

Winglee M, Valliant R, Scheuren F (2005) A case study in record linkage. Surv Methodol 31(1):3–12

Winkler WE (1990) String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. In: JSM proceedings, Survey research methods section. American Statistical Association, Alexandria, VA, pp 354–359 Retrieved from http://ww2.amstat.org/sections/srms/Proceedings/. Accessed 2018

Natalie Shlomo (BSc, Mathematics and Statistics, Hebrew University; MA, Statistics, Hebrew University; PhD, Statistics, Hebrew University) is Professor of Social Statistics in the School of Social Sciences at the University of Manchester. Her area of interest is in survey statistics covering survey design and estimation, record linkage, statistical disclosure control, statistical data editing and imputation, non-response analysis and adjustments, adaptive survey designs, quality indicators for survey representativeness and small area estimation.

Privacy in Microdata Release: Challenges, Techniques, and Approaches

Giovanni Livraga

1 Introduction

We live in a society that relies more and more on the availability of data to make knowledge-based decisions (Livraga, 2015). The benefits that can be driven by data sharing and dissemination have been widely recognized for a long time now (Foresti, 2011; Livraga, 2015), and are visible to everybody: for instance, medical research is a simple example of a field that, leveraging analysis of real clinical trials made available by hospitals, can improve the life quality of individuals. At the same time, many laws and regulations have recognized that privacy is a primary right of citizens, acknowledging the principle that sensitive information (e.g., personal information that refers to an individual) must be protected from improper disclosure. To resolve the tension between the (equally strong) needs for data privacy and availability, the scientific community has been devoting major efforts for decades to investigating models and approaches that can allow a data owner to release a data collection guaranteeing that sensitive information be properly protected, while still allowing useful analysis to be performed (Bezzi et al., 2012; De Capitani di Vimercati et al., 2011b).

In the past, data were typically released in the form of aggregate statistics (*macrodata*): while providing a first layer of protection to the individuals to whom the statistics pertain, as no specific data of single respondents (i.e., the individuals to whom data items refer) are (apparently) disclosed (De Capitani di Vimercati et al., 2011a), releasing precomputed statistics inevitably limits the analysis that a recipient can do. To provide recipients with greater flexibility in performing analysis, many situations require the release of detailed data, called *microdata*.

G. Livraga (✉)
Dipartimento di Informatica, Università degli Studi di Milano, Crema, Italy
e-mail: giovanni.livraga@unimi.it

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_5

Indeed, since analyses are not precomputed, more freedom is left to the final recipients. The downside, however, comes in terms of major privacy concerns, as microdata can include sensitive information precisely related to individuals.

As will be illustrated in this chapter, the first attempts towards the development of microdata protection approaches pursued what today are typically called *syntactic* privacy guarantees (Ciriani et al., 2007a; Clifton and Tassa, 2013; De Capitani di Vimercati et al., 2012). Traditional protection approaches (e.g., k-anonymity (Samarati, 2001) and its variations) operate by removing and/or generalizing (i.e., making less precise/more general) all information that can identify a respondent, so that each respondent is hidden in a group of individuals sharing the same identifying information. In this way, it is not possible to precisely link an individual to her (sensitive) information. Existing solutions following this approach can be used to protect respondents' identities as well as their sensitive information (Livraga, 2015), also in emerging scenarios (De Capitani di Vimercati et al., 2015b). Alternative approaches based on the notion of differential privacy (Dwork, 2006) have then been proposed. Trying to pursue a relaxed and microdata-adapted version of a well-known definition of privacy by Dalenius (1977), that anything that can be learned about a respondent from a statistical database should be learnable without access to the database, differential privacy aims at ensuring that the inclusion in a dataset of the information of an individual does not significantly alter the outcome of analysis of the dataset. To achieve its privacy goal, differential privacy typically relies on controlled noise addition, thus perturbing the data to be released (in contrast to k-anonymity-like solutions that, operating through generalization, guarantee data truthfulness). There has been a major debate in the scientific community regarding which approach (syntactic techniques versus differential privacy) is the "correct" one (Clifton and Tassa, 2013; Kifer and Machanavajjhala, 2011), and recent studies have pointed out that, while they pursue different privacy goals through different protection techniques, both approaches are successfully applicable to different scenarios, and there is room for both of them (Clifton and Tassa, 2013; Li et al., 2012a), possibly jointly adopted (Soria-Comas et al., 2014). Both the approaches have in fact been used in different application scenarios, ranging from the protection of location data (e.g., Peng et al. 2016; Xiao and Xiong 2015), to privacy-preserving data mining (e.g., Ciriani et al. 2008; Li et al. 2012c), and to the private analysis of social network data (e.g., Tai et al. 2014; Wang et al. 2016), just to name a few.

The goal of this chapter is to illustrate some of the best-known protection techniques and approaches that can be used to ensure microdata privacy. The remainder of this chapter is organized as follows. Section 2 presents the basic concepts behind the problem of microdata protection, illustrating possible privacy risks and available protection techniques. Section 3 discusses some well-known protection approaches. Section 4 illustrates some extensions of the traditional approaches, proposed to relax or remove some assumptions for use in advanced scenarios, with a specific focus on the problem of protecting microdata coming from multiple sources. Finally, Sect. 5 concludes the chapter.

2 Microdata Protection: Basic Concepts

This section illustrates the key concepts behind the problem of protecting microdata privacy. It discusses firstly some privacy issues that can arise in microdata release (Sect. 2.1), and secondly the protection techniques that have been proposed by the research community to protect microdata (Sect. 2.2).

2.1 Microdata Privacy

Microdata can be represented as relational tables including a set of tuples, related to a set of individuals (called *respondents*), and defined over a set of attributes. Traditional data protection approaches classify attributes in a microdata table depending on their identifying ability and sensitivity, as follows (Ciriani et al., 2007a).[1]

- *Identifiers*: attributes that uniquely identify a respondent (e.g., `Name` and `SSN`).
- *Quasi-identifiers* (QI): attributes that, in combination, can be linked to external information to reidentify (all or some of) the respondents to whom information refers, or to reduce the uncertainty over their identities (e.g., `DoB`, `Sex`, and `ZIP`).
- *Sensitive attributes*: attributes that represent information that should be kept confidential (e.g., `Disease`).

The first step in protecting a microdata table to be released is to remove (e.g., by deleting or encrypting) all identifiers from the table. This process, usually referred to as *de-identification*, is unfortunately not sufficient to effectively ensure the *anonymity* of the data, due to the presence of QI attributes (e.g., 63% of the entire US population in the US 2000 Census was *uniquely identifiable* by the combination of their gender, ZIP code, and full date of birth (Golle et al., 2006)). To illustrate, consider the de-identified version of a microdata table including information on a set of hospitalized patients in Fig. 1a. Figure 1b illustrates a sample excerpt of a (fictitious) publicly available voter list for the municipality of New York City. Attributes `DoB`, `Sex`, and `ZIP` can be used to link the two tables, allowing the re-identification (with either full confidence or a certain probability) of some of the de-identified respondents in Fig. 1a. For instance, the de-identified microdata include only one female respondent, born in 1958/12/11 and living in the 10180 area (tuple 11). If this combination of QI values is unique in the external world as well, the voter list can be exploited to uniquely reidentify the eleventh tuple with respondent *Kathy Doe*, also disclosing the fact that she has been hospitalized for

[1] In this chapter, `SSN`, `DoB`, and `ZIP` are attributes representing Social Security Numbers (the de facto US identification number for taxation and other purposes), dates of birth, and ZIP codes (US postal codes).

SSN	Name	DoB	Sex	ZIP	Disease
		1960/05/02	F	10041	stroke
		1960/05/20	M	10032	dyspepsia
		1960/05/12	M	10037	achlorhydria
		1960/05/05	F	10044	epilepsy
		1955/09/01	M	10043	helicobacter
		1955/09/02	M	10042	helicobacter
		1955/09/10	F	10039	helicobacter
		1955/09/20	F	10030	helicobacter
		1955/12/07	M	10030	dermatitis
		1955/12/05	M	10031	retinitis
		1958/12/11	*F*	*10180*	epilepsy
		1955/12/25	F	10042	dermatitis
		1955/12/30	F	10045	gastritis
		1960/04/02	F	10036	stroke
		1960/04/05	F	10034	labyrinthitis
		1960/04/10	M	10047	gastritis
		1960/04/30	M	10048	dyspepsia

(a)

Name	Address	City	ZIP	DoB	Sex	Education
...	...	...	...	...	...	...
Kathy Doe	300 Main St.	New York City	*10180*	*58/12/11*	*female*	secondary
...	...	...	...	...	...	...

(b)

Fig. 1 An example of a de-identified microdata table (**a**) and of a publicly available non-de-identified dataset (**b**)

epilepsy. Given that tremendous amounts of data are generated and shared every day, the availability of non-de-identified datasets that can be used for linking is a realistic threat. Unfortunately, unlike direct identifiers, QI cannot be easily removed to protect privacy, since QI attributes can represent a large portion of the attributes in the table, and their complete removal would reduce the utility of the anonymized data too much (e.g., removing also the QI from the de-identified microdata in Fig. 1a would leave only a list of diseases, most probably of limited interest to the final recipients).

Given a de-identified microdata table, two different kinds of improper disclosure can occur, as follows (Federal Committee on Statistical Methodology, 2005).

- *Identity disclosure*, occurring whenever the identity of a respondent can be somehow determined and associated with a (de-identified) tuple in the released microdata table.

- *Attribute disclosure*, occurring when a (sensitive) attribute value can be associated with an individual (without necessarily being able to link the value to a specific tuple).

2.2 Protection Techniques

Various *microdata protection techniques* have recently been proposed by the scientific community (Ciriani et al., 2007b; Federal Committee on Statistical Methodology, 2005). An initial distinction can be made between *masking techniques* and *synthetic data generation techniques*: while these latter aim to release a new, synthetic dataset that preserves some statistical properties of the original data, masking techniques operate directly on the original microdata, to sanitize them before release, and can be classified as follows.

- *Non-perturbative techniques* do not directly modify the original data, but remove details from the microdata table: they sacrifice data completeness by releasing possibly imprecise and/or incomplete data to preserve data truthfulness. Examples of non-perturbative techniques include *suppression*, *generalization*, and *bucketization*. Suppression selectively removes information from the microdata table. Generalization, possibly based on ad hoc generalization hierarchies, selectively replaces the content of some cells in the microdata table (e.g., a complete date of birth) with more general values (e.g., year of birth). Bucketization operates on sets of attributes whose joint visibility should be prevented (e.g., the name and the disease of a patient), and operates by first partitioning tuples in buckets and attributes in groups, and then shuffling the semi-tuples within buckets so as to break their correspondence (De Capitani di Vimercati et al., 2015a, 2010; Li et al., 2012b; Xiao and Tao, 2006).
- *Perturbative techniques* distort the microdata table to be released by modifying its informative content, hence sacrificing data truthfulness. Examples of perturbative techniques include *noise addition* and *microaggregation*. Noise addition intuitively adds controlled noise to the original data collection. Protection is provided by the fact that some values (or combinations among them) included in the released table might not correspond to real ones, and vice versa. Microaggregation (originally proposed for continuous numerical data and then extended also to categorical data (Torra, 2004)) selectively replaces original tuples with new ones. It operates by first clustering the tuples in the original microdata table in groups of a certain cardinality in such a way that tuples in the same cluster are similar to each other, and then by replacing the tuples in a cluster with a representative one computed through an aggregation operator (e.g., mean or median).

The protection techniques illustrated above can be adopted to effectively protect the confidentiality of a microdata collection to be released. Given a data collection to be protected and released, some key questions then need to be answered: what technique should be used? Should a combination of techniques be preferred to a single one? To which portion of the data (e.g., the entire table, a subset of tuples, and a subset of attributes) should the technique be applied? Whatever the answers to these questions, an important observation is that all microdata protection techniques cause an inevitable information loss: non-perturbative techniques produce datasets that are not as complete or as precise as the originals, and perturbative techniques produce datasets that are distorted. For these reasons, the scientific community has recently developed protection approaches that, given a privacy requirement to be satisfied (e.g., the protection of the identities of the microdata respondents), rely on a controlled adoption of some of these microdata protection techniques to protect privacy while limiting information loss, as illustrated in the remainder of this chapter.

3 Microdata Protection Approaches

This section illustrates the most important protection approaches that have driven research in microdata protection in the past couple of decades, together with the privacy requirements they pursue and the microdata protection techniques (see Sect. 2) that are typically adopted for their enforcement.

3.1 k-Anonymity

The first and pioneering approach for protecting microdata against identity disclosure is represented by k-anonymity (Samarati, 2001), enforcing a protection requirement typically applied by statistical agencies that demands that any released information be *indistinguishably related* to no less than a certain number k of respondents. Following the assumption that re-identification of de-identified microdata takes advantage of QI attributes, such general requirement is translated into the k-anonymity requirement: each release of data must be such that every *combination of values of the QI* can be indistinctly matched to *at least k respondents* (Samarati, 2001). A microdata table satisfies the k-anonymity requirement iff each tuple cannot be related to less than k individuals in the population, and vice versa (i.e., each individual in the population cannot be related to less than k tuples in the table). These two conditions hold since the original definition of k-anonymity assumes that each respondent is represented by at most one tuple in the released table and vice versa (i.e., each tuple includes information related to one respondent only).

Verifying the satisfaction of the k-anonymity requirement would require knowledge of *all* existing external sources of information that an adversary might use for the linking attack. This assumption is indeed unrealistic in practice, and therefore k-anonymity takes the safe approach of requiring that each respondent be indistinguishable from at least $k - 1$ other respondents in the released microdata.

A table is therefore said to be *k*-anonymous if each combination of values of the QI appears in it with either zero or at least *k* occurrences. For instance, the table in Fig. 1a is 1-anonymous if we assume the QI to be composed of `DoB`, `Sex`, and `ZIP`, since at least one combination of their values (i.e., ⟨1958/12/11, F, 10180⟩) appears only once in the table (i.e., in the eleventh tuple). Since each combination of QI values is shared by at least *k* different tuples in the microdata table, each respondent cannot be associated with fewer than *k* tuples in the released table and vice versa, also satisfying the original *k*-anonymity requirement (being the definition of a *k*-anonymous table a sufficient, though not necessary, condition for the satisfaction of the *k*-anonymity requirement).

Traditional approaches to enforcing *k*-anonymity operate on QI attributes by modifying their values in the microdata to be released, while leaving sensitive and nonsensitive attributes as they are (recall that direct identifiers are removed from the microdata as the first step). Among the possible data protection techniques that might be enforced on the QI, *k*-anonymity typically relies on the combined adoption of *generalization* and *suppression*, which have the advantage of preserving data truthfulness when compared to perturbative techniques (e.g., noise addition; see Sect. 2.2). Suppression is used to couple generalization, as it can help in reducing the amount of generalization that has to be enforced to achieve *k*-anonymity; in this way, it is possible to produce more precise (though incomplete) tables. The intuitive rationale is that, if a microdata table includes a limited number of outliers (i.e., QI values with less than *k* occurrences) that would force a large amount of generalization to satisfy *k*-anonymity, these outliers could be more conveniently removed from the table, improving the quality of the released data.

Generalization and suppression can be applied at various granularity levels (i.e., generalization at the cell and attribute levels, and suppression at the cell, attribute, and tuple levels), and the combined use of generalization and suppression at different granularity levels produces different classes of approaches to enforcing *k*-anonymity (Ciriani et al., 2007a). The majority of the approaches available in the literature adopt attribute-level generalization and tuple-level suppression (Bayardo and Agrawal, 2005; LeFevre et al., 2005; Samarati, 2001). Figure 2 illustrates a 4-anonymous table obtained from the microdata in Fig. 1a through attribute-level generalization (`DoB`, `Sex`, and `ZIP` have been generalized by removing the day of birth, sex, and the last two digits of the ZIP code, respectively) and tuple-level suppression (the 11th tuple related to *Kathy* has been suppressed). Cell-level generalization has also been investigated as an approach to producing *k*-anonymous tables (LeFevre et al., 2006). To reduce the inevitable information loss (the original microdata informative content is either reduced in detail or removed), it is necessary to compute an optimal *k*-anonymization minimizing generalization and suppression, which has been shown to be an NP-hard problem (Ciriani et al., 2007a), and both exact and heuristic algorithms have been proposed.

As a last remark on *k*-anonymity, it should be noted that some recent approaches have been proposed to obtain *k*-anonymity through microaggregation (see Sect. 2.2) (Domingo-Ferrer and Torra, 2005; Soria-Comas et al., 2014). To this end, the QI undergoes microaggregation, so that each combination of

Fig. 2 An example of 4-anonymous table

SSN	Name	DoB	Sex	ZIP	Disease
		1960/05	*	100**	stroke
		1960/05	*	100**	dyspepsia
		1960/05	*	100**	achlorhydria
		1960/05	*	100**	epilepsy
		1955/09	*	100**	helicobacter
		1955/09	*	100**	helicobacter
		1955/09	*	100**	helicobacter
		1955/09	*	100**	helicobacter
		1955/12	*	100**	dermatitis
		1955/12	*	100**	retinitis
		1955/12	*	100**	dermatitis
		1955/12	*	100**	gastritis
		1960/04	*	100**	stroke
		1960/04	*	100**	labyrinthitis
		1960/04	*	100**	gastritis
		1960/04	*	100**	dyspepsia

Fig. 3 An example of a microdata table (**a**) and of a 3-anonymous version of it (**b**) obtained by adopting microaggregation

Age	Disease
20	flu
25	gastritis
30	dermatitis
35	stroke
40	dyspepsia
45	asthma

(a)

Age	Disease
25	flu
25	gastritis
25	dermatitis
40	stroke
40	dyspepsia
40	asthma

(b)

QI values in the original microdata table is replaced with a microaggregated version. Figure 3b illustrates a 3-anonymous version of the microdata in Fig. 3a obtained through microaggregation, assuming `Age` to be the QI, and `Disease` the sensitive attribute. Note that, being microaggregation a perturbative protection technique, k-anonymous tables computed adopting this approach do not preserve data truthfulness.

3.2 ℓ-Diversity and t-Closeness

While k-anonymity represents an effective solution to protect respondent identities, it does not protect against attribute disclosure (Samarati, 2001). A k-anonymous table can in fact still be vulnerable to attacks allowing a recipient to determine with non-negligible probability the sensitive information of a respondent, as follows (Machanavajjhala et al., 2007; Samarati, 2001).

- *Homogeneity attack*. A homogeneity attack occurs when all the tuples in an equivalence class (i.e., the set of tuples with the same value for the QI) in a k-anonymous table assume the same value for the sensitive attribute. If a data recipient knows the QI value of a target individual x, she can identify the equivalence class representing x, and then discover the value of x's sensitive attribute. For instance, consider the 4-anonymous table in Fig. 2 and suppose that a recipient knows that *Gloria* is a female living in the 10039 area and born on 1955/09/10. Since all the tuples in the equivalence class with QI value equal to $\langle 1955/09, *, 100**\rangle$ assume value *helicobacter* for attribute `Disease`, the recipient can infer that Gloria suffers from a *helicobacter* infection.
- *External knowledge attack*. The external knowledge attack occurs when the data recipient possesses some additional knowledge (not included in the k-anonymous table) about a target respondent x, and can use it to reduce the uncertainty about the value of x's sensitive attribute. For instance, consider the 4-anonymous table in Fig. 2 and suppose that a recipient knows that a neighbor, *Mina*, is a female living in the 10045 area and born on 1955/12/30. Observing the 4-anonymous table, the recipient can infer only that the neighbor suffers from *dermatitis*, *retinitis*, or *gastritis*. Suppose now that the recipient sees *Mina* tanning without screens at the park every day: due to this external information, the recipient can exclude the likelihood that *Mina* suffers from *dermatitis* or *retinitis*, and infer that she suffers from *gastritis*.

The original definition of k-anonymity has been extended to ℓ-diversity to counteract these two forms of attack. The idea behind ℓ-diversity is to take into account the values of the sensitive attributes when clustering the original tuples, so that at least ℓ *well-represented* values for the sensitive attribute are included in each equivalence class (Machanavajjhala et al., 2007). While several definitions for "well-represented" values have been proposed, the simplest formulation of ℓ-diversity requires that each equivalence class be associated with at least ℓ different values for the sensitive attribute. For instance, consider the 4-anonymous and 3-diverse table in Fig. 4 and suppose that a recipient knows that a neighbor, *Mina*, a female living in the 10045 area and born on 1955/12/30, tans every day at the park (see example above). The recipient can now only exclude value *dermatitis*, but she cannot be sure about whether *Mina* suffers from *gastritis* or a *helicobacter* infection.

Computing an ℓ-diverse table minimizing the loss of information caused by generalization and suppression is computationally hard. However, since ℓ-diversity basically requires computing a k-anonymous table (with additional constraints on the sensitive values), any algorithm proposed for computing a k-anonymous table that minimizes loss of information can be adapted to also guarantee ℓ-diversity, simply by controlling whether or not the condition on the diversity of the sensitive attribute values is satisfied by all the equivalence classes (Machanavajjhala et al., 2007). As a last remark on ℓ-diversity, it might be possible to obtain ℓ-diverse tables by departing from generalization and adopting instead a bucketization-based approach (see Sect. 2.2), for instance, by adopting the Anatomy approach (Xiao and

Fig. 4 An example of 4-anonymous and 3-diverse table

SSN	Name	DoB	Sex	ZIP	Disease
		1955	M	100**	helicobacter
		1955	M	100**	helicobacter
		1955	M	100**	dermatitis
		1955	M	100**	retinitis
		1960	F	100**	stroke
		1960	F	100**	epilepsy
		1960	F	100**	stroke
		1960	F	100**	labyrinthitis
		1955	F	100**	helicobacter
		1955	F	100**	helicobacter
		1955	F	100**	dermatitis
		1955	F	100**	gastritis
		1960	M	100**	dyspepsia
		1960	M	100**	achlorhydria
		1960	M	100**	gastritis
		1960	M	100**	dyspepsia

Tao, 2006), or other (possibly more general) techniques (Ciriani et al., 2012; De Capitani di Vimercati et al., 2014, 2015a, 2010).

Although ℓ-diversity represents a first step in counteracting attribute disclosure, an ℓ-diverse table might still be vulnerable to information leakage caused by *skewness attacks* (where significant differences can be seen in the frequency distribution of the sensitive values within an equivalence class with respect to that of the same values in the overall population), and *similarity attacks* (where the ℓ sensitive values of the tuples in an equivalence class are semantically similar, although syntactically different) (Li et al., 2007). To counteract these two disclosure risks, it is possible to rely on the definition of t-closeness (Li et al., 2007), requiring that the frequency distribution of the sensitive values in each equivalence class be close (i.e., with distance smaller than a fixed threshold t) to that in the released microdata table.

3.3 Differential Privacy

Differential privacy (DP) is a recent privacy definition that departs from the guarantees and enforcement techniques characterizing k-anonymity and its extensions, and aims to guarantee that the release of a dataset does not disclose sensitive information about *any* individual, who may or may not be represented therein (Dwork, 2006). DP aims at releasing a dataset permitting the disclosure of *properties* about the population as a whole (rather than the microdata themselves), while protecting the privacy of single individuals. The privacy guarantee provided by DP relies on ensuring that the probability of a recipient correctly inferring the sensitive value of a target respondent x be not affected by the presence or absence of x's tuple in the released dataset.

DP can be adopted either to respond to queries (*interactive scenario*) issued against a microdata table or to produce a sanitized dataset to be released (*noninteractive scenario*). In the interactive scenario, DP is ensured by adding *random noise* to the query results evaluated on the original dataset (Dwork et al., 2006), sacrificing data truthfulness. Unfortunately, the interactive scenario limits the analysis that the recipient can perform, as it allows only a limited number of queries to be answered (Soria-Comas et al., 2014). In the noninteractive scenario, a dataset is produced and released, typically based on the evaluation of histogram queries (i.e., counting the number of records having a given value). To reduce information leakage, these counts are computed through a DP mechanism.

Unlike k-anonymity and its variations, which guarantee a certain degree of privacy to the microdata to be released, DP aims to guarantee that the *release mechanism* $\mathcal{K}$ (e.g., the algorithm adopted to compute the data to be released, whether query answers in the interactive scenario or sanitized counts in the noninteractive scenario) is safe with respect to privacy breaches. A dataset to be released satisfies DP if the removal/insertion of one tuple from/to the dataset does not significantly affect the result of the evaluation of $\mathcal{K}$. In this way, the protection offered by DP lies in the fact that the impact that a respondent has on the outcome of a certain analysis (or on the generation of the sanitized dataset) remains negligible. In fact, DP guarantees that the probability of observing a result for the evaluation of $\mathcal{K}$ over T is close to the probability of observing that result for the evaluation of $\mathcal{K}$ over a dataset T' differing from T for a tuple only.

DP offers strong privacy guarantees at the price of imposing strict conditions on what kind of, and how, data can be released (Clifton and Tassa, 2013). In addition, the amount of noise that needs to be adopted can significantly distort the released data (Clifton and Tassa, 2013; Fredrikson et al., 2014; Soria-Comas et al., 2014), thus limiting in practice their utility for final recipients. Some relaxations of DP have therefore been proposed (e.g., Dwork and Smith 2009; Mironov et al. 2009), possibly applicable to specific real-world scenarios (e.g., Hong et al. 2015), with the aim of finding a reasonable tradeoff between privacy protection and data utility.

It is interesting to note that a recent approach has been proposed using k-anonymity and DP approaches together, with the aim of reducing the amount of noise needed to ensure DP (Soria-Comas et al., 2014). The proposal builds on the observation that, given a microdata table T and a query q for which the outputs are required to be differentially private, if the query is run on a microaggregation-based (see Sect. 3.1) k-anonymous version T_k of T, the amount of noise to be added to the output of q for achieving DP is greatly reduced (compared with the noise that would be needed if q were run on the original T). To this end, microaggregation should be performed carefully so that it can be considered *insensitive* to the input data (i.e., for any pair of datasets T and T' differing by one tuple, given the clusters $\{c_1, \ldots, c_n\}$ produced by the microaggregation over T and the clusters $\{c'_1, \ldots, c'_n\}$ produced by the microaggregation over T', each pair of corresponding clusters differs in at most one tuple). This is a key property required for the microaggregation to succeed in reducing the noise that will then be employed to ensure DP, as it reduces the sensitivity of the query to be executed (Soria-Comas et al., 2014) and hence the result distortion. This approach can also be used in the noninteractive scenario. To this end, a k-anonymous version T_k of T is first built through an insensitive microaggregation. The differentially private dataset T_{DP} is then built by collating the n differentially private answers to a set of n queries (with n the number of tuples in T_k), where the ith query ($i = 1, \ldots, n$) aims at retrieving the ith tuple in T_k.

4 Extensions for Advanced Scenarios

The traditional microdata protection approaches in the literature (see Sect. 3) are built on specific assumptions that can limit their applicability to certain scenarios. For instance, they assume the data to be released in a single table, completely available for anonymization before release, and never republished. However, it may happen that data are either republished over time or continuously generated, as in the case with data streams: recent proposals (e.g., Fung et al. 2008; Loukides et al. 2013; Shmueli and Tassa 2015; Shmueli et al. 2012; Tai et al. 2014; Xiao and Tao 2007) have extended traditional approaches to deal with these scenarios.

One of the assumptions on which the original formulations of k-anonymity, DP, and their extensions were based is that the microdata to be anonymized are stored in a single table. This assumption represents a limitation in many real-world scenarios, in which the information that needs to be released can be spread across various datasets, and where the privacy goal is that *all* released information be effectively protected. There are two naive approaches that one might think of adopting: join-and-anonymize and anonymize-and-join. The first approach, in which all tables to be released are first joined in a universal relation that is then anonymized by adopting one of the traditional approaches, might not work whenever there is no single subject authorized to see and join all original relations, which might be owned by different authorities. The second approach (i.e., first anonymize each table singularly taken and then release the join among the sanitized versions of all tables) does not guarantee appropriate protection: for instance, if a QI is spread across multiple tables, it could not be effectively anonymized by looking at each relation individually. The scientific community has recently started looking at this problem, and some solutions have been proposed (typically extending k-anonymity and its variations) to address the multiple tables scenario.

A first distinction has to be made depending on whether the multiple tables to be released belong to the same authority (e.g., different relations of a single database) that therefore has a complete view over them, or the tables belong to different authorities, where no subject in the picture has a global view of the entire informative content that needs to be released. In the first scenario, a careful join-and-anonymize approach might do. However, the anonymization has to be performed with extreme care to avoid vulnerability to privacy breaches. For instance, assume n relations, owned by the same authority, to be released together provided that k-

anonymity is satisfied by their join. When computing the join among the n relations, it might be possible that the k-anonymity assumption of one respondent being represented by a single tuple is not satisfied (as different tuples could be related to the same respondent). The risk here is that (some of) the different tuples related to the same individual are "anonymized together": hence, an equivalence class of size k might refer to less than k respondents, violating their privacy despite the relation being apparently k-anonymous. To overcome this issue, MultiR k-anonymity (Nergiz et al., 2007) has been proposed to extend the definition of k-anonymity and ℓ-diversity to multiple relations belonging to a snowflake database schema.

When the relations to be anonymized belong to different authorities, it is clearly not possible to join them beforehand. One might think to first anonymize each relation individually and then join the obtained results on the (anonymized) QI. Unfortunately, this strategy is not trivial: besides possibly exploding in size, the joined tuples could not be used for meaningful analysis, as many tuples in the join would be incorrect (joining over the anonymized QI would join more tuples than using the original values). Some approaches have recently been proposed to address this issue. For instance, distributed k-anonymity (DkA (Jiang and Clifton, 2006)) proposes a distributed framework for achieving k-anonymity. The applicability of this approach is limited to two relations (defined as two views over a global data collection), which can be correctly joined through a 1:1 join on a common key. The framework builds a k-anonymous join of the two datasets, without disclosing any information from one site to the other. In a nutshell, the approach works iteratively in three steps: (1) each data holder produces a k-anonymous version of her own dataset; (2) each data holder checks whether or not joining the obtained k-anonymous datasets would maintain global k-anonymity; and (3) if so, join and release, otherwise go back to step 1 and further generalize the original data. Checking the global anonymity (step 2) is a critical task, as it requires the two parties to exchange their anonymized tables. To avoid information leakage, encryption is adopted and, in this regard, the price to be paid for this approach is in terms of the required encryption and decryption overhead (Jiang and Clifton, 2006; Mohammed et al., 2011). Recent efforts that have recently been devoted to enforce DP in a multi-relational setting (Mohammed et al., 2014) (also focusing on two relations only) should also be highlighted. The solution in Mohammed et al. (2011) instead does not pose assumptions on the number of relations to be joined but requires active cooperation among the parties holding the relations to achieve k-anonymity. In addition, the approach in Mohammed et al. (2011) can be successfully extended to provide privacy beyond k-anonymity (e.g., by ensuring ℓ-diversity). Finally, it should be noted that specific approaches have also been proposed to protect different tables that need to be *sequentially* released (Wang and Fung, 2006).

5 Conclusions

This chapter has addressed the problem of protecting privacy in microdata release. After a discussion of the privacy risks that can arise when microdata need to be shared or disseminated, some of the best-known microdata protection techniques and approaches developed by the scientific community have been illustrated. Some recent extensions of traditional approaches, proposed to fit advanced scenarios, have also been highlighted.

Acknowledgements This paper is based on joint work with Sabrina De Capitani di Vimercati, Sara Foresti, and Pierangela Samarati, whom the author would like to thank. This work was supported in part by the European Commission through the Seventh Framework Programme under grant agreement 312797 (ABC4EU) and through the Horizon 2020 programme under grant agreement 644579 (ESCUDO-CLOUD).

References

Bayardo RJ, Agrawal R (2005) Data privacy through optimal *k*-anonymization. In: Proceedings of ICDE 2005, Tokyo, April 2005

Bezzi M, De Capitani di Vimercati S, Foresti S, Livraga G, Samarati P, Sassi R (2012) Modeling and preventing inferences from sensitive value distributions in data release. J Comput Secur 20(4):393–436

Ciriani V, De Capitani di Vimercati S, Foresti S, Samarati P (2007) *k*-anonymity. In: Yu T, Jajodia S (eds) Secure data management in decentralized systems. Springer, Berlin

Ciriani V, De Capitani di Vimercati S, Foresti S, Samarati P (2007) Microdata protection. In: Yu T, Jajodia S (eds) Secure data management in decentralized systems. Springer, Berlin

Ciriani V, De Capitani di Vimercati S, Foresti S, Samarati P (2008) *k*-Anonymous data mining: a survey. In: Aggarwal C, Yu P (eds) Privacy-preserving data mining: models and algorithms. Springer, Berlin

Ciriani V, De Capitani di Vimercati S, Foresti S, Livraga G, Samarati P (2012) An OBDD approach to enforce confidentiality and visibility constraints in data publishing. J Comput Secur 20(5):463–508

Clifton C, Tassa T (2013) On syntactic anonymity and differential privacy. Trans Data Priv 6(2):161–183

Dalenius T (1977) Towards a methodology for statistical disclosure control. Statistik Tidskrift 15:429–444

De Capitani di Vimercati S, Foresti S, Jajodia S, Paraboschi S, Samarati P (2010) Fragments and loose associations: respecting privacy in data publishing. Proc VLDB Endow 3(1):1370–1381

De Capitani di Vimercati S, Foresti S, Livraga G, Samarati P (2011) Anonymization of statistical data. Inform Technol 53(1):18–25

De Capitani di Vimercati S, Foresti S, Livraga G, Samarati P (2011) Protecting privacy in data release. In: Aldini A, Gorrieri R (eds) Foundations of security analysis and design VI. Springer, Berlin

De Capitani di Vimercati S, Foresti S, Livraga G, Samarati P (2012) Data privacy: definitions and techniques. Int J Uncertainty Fuzziness Knowl Based Syst 20(6):793–817

De Capitani di Vimercati S, Foresti S, Jajodia S, Livraga G, Paraboschi S, Samarati P (2014) Fragmentation in presence of data dependencies. IEEE Trans Dependable Secure Comput 11(6):510–523

De Capitani di Vimercati S, Foresti S, Jajodia S, Livraga G, Paraboschi S, Samarati P (2015) Loose associations to increase utility in data publishing. J Comput Secur 23(1):59–88

De Capitani di Vimercati S, Foresti S, Livraga G, Paraboschi S, Samarati P (2015) Privacy in pervasive systems: social and legal aspects and technical solutions. In: Colace F, Santo MD, Moscato V, Picariello A, Schreiber F, Tanca L (eds) Data management in pervasive systems. Springer, Berlin

Domingo-Ferrer J, Torra V (2005) Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Min Knowl Disc 11(2):195–212

Dwork C (2006) Differential privacy. In: Proceedings of ICALP 2006, Venice, July 2006

Dwork C, Smith A (2009) Differential privacy for statistics: what we know and what we want to learn. J Priv Confid 1(2):135–154

Dwork C, Mcsherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: Proceedings of TCC 2006, New York, NY, March 2006

Federal Committee on Statistical Methodology (2005) Statistical policy working paper 22 (Second Version). Report on statistical disclosure limitation methodology, December 2005

Foresti S (2011) Preserving privacy in data outsourcing. Springer, Berlin

Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T (2014) Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In: Proceedings of the 23rd USENIX security symposium, San Diego, August 2014

Fung BCM, Wang K, Fu AWC, Pei J (2008) Anonymity for continuous data publishing. In: Proceedings of EDBT 2008, Nantes, March 2008

Golle P (2006) Revisiting the uniqueness of simple demographics in the US population. In: Proceedings of WPES 2006, Alexandria, October 2006

Hong Y, Vaidya J, Lu H, Karras P, Goel S (2015) Collaborative search log sanitization: toward differential privacy and boosted utility. IEEE Trans Dependable Secure Comput 12(5):504–518

Jiang W, Clifton C (2006) A secure distributed framework for achieving k-anonymity. VLDB J 15(4):316–333

Kifer D, Machanavajjhala A (2011) No free lunch in data privacy. In: Proceedings of SIGMOD 2011, Athens, June 2011

LeFevre K, DeWitt D, Ramakrishnan R (2005) Incognito: efficient full-domain k-anonymity. In: Proceedings of SIGMOD 2005, Baltimore, June 2005

LeFevre K, DeWitt D, Ramakrishnan R (2006) Mondrian multidimensional k-anonymity. In: Proceedings of ICDE 2006, Atlanta, April 2006

Li N, Li T, Venkatasubramanian S (2007) t-closeness: privacy beyond k-anonymity and ℓ-diversity. In: Proceedings of ICDE 2007, Istanbul

Li N, Qardaji W, Su D (2012) On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In: Proceedings of ASIACCS 2012, Seoul, May 2012

Li T, Li N, Zhang J, Molloy I (2012) Slicing: a new approach for privacy preserving data publishing. IEEE Trans Knowl Data Eng 24(3):561–574

Li Y, Chen M, Li Q, Zhang W (2012) Enabling multilevel trust in privacy preserving data mining. IEEE Trans Knowl Data Eng 24(9):1598–1612

Livraga G (2015) Protecting privacy in data release. Springer, Berlin

Loukides G, Gkoulalas-Divanis A, Shao J (2013) Efficient and flexible anonymization of transaction data. Knowl Inform Syst. 36(1):153–210

Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007) ℓ-Diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov from Data 1(1):3:1–3:52

Mironov I, Pandey O, Reingold O, Vadhan S (2009) Computational differential privacy. In: Proceedings of CRYPTO 2009, Santa Barbara, August 2009

Mohammed N, Fung BC, Debbabi M (2011) Anonymity meets game theory: secure data integration with malicious participants. VLDB J 20(4):567–588

Mohammed N, Alhadidi D, Fung BC, Debbabi M (2014) Secure two-party differentially private data release for vertically partitioned data. IEEE Trans Dependable Secure Comput 11(1): 59–71

Nergiz M, Clifton C, Nergiz A (2007) Multirelational k-anonymity. In: Proceedings of ICDE 2007, Istanbul

Peng T, Liu Q, Meng D, Wang G (2017) Collaborative trajectory privacy preserving scheme in location-based services. Inform Sci 387:165–179. Available online

Samarati P (2001) Protecting respondents' identities in microdata release. IEEE Trans Knowl Data Eng 13(6):1010–1027

Shmueli E, Tassa T (2015) Privacy by diversity in sequential releases of databases. Inform Sci 298:344–372

Shmueli E, Tassa T, Wasserstein R, Shapira B, Rokach L (2012) Limiting disclosure of sensitive data in sequential releases of databases. Inform Sci 191:98–127

Soria-Comas J, Domingo-Ferrer J, Sánchez D, Martínez S (2014) Enhancing data utility in differential privacy via microaggregation-based k-anonymity. VLDB J 23(5):771–794

Tai CH, Tseng PJ, Yu PS, Chen MS (2014) Identity protection in sequential releases of dynamic networks. IEEE Trans Knowl Data Eng 26(3):635–651

Torra V (2004) Microaggregation for categorical variables: a median based approach. In: Proceedings of PSD 2004, Barcelona, June 2004

Wang K, Fung B (2006) Anonymizing sequential releases. In: Proceedings of KDD 2006, Philadelphia, August 2006

Wang Q, Zhang Y, Lu X, Wang Z, Qin Z, Ren K (2016) Real-time and spatio-temporal crowdsourced social network data publishing with differential privacy. IEEE Trans Dependable Secure Comput (in press)

Xiao X, Tao Y (2006) Anatomy: simple and effective privacy preservation. In: Proceedings of VLDB 2006, Seoul, September 2006

Xiao X, Tao Y (2007) m-Invariance: towards privacy preserving re-publication of dynamic datasets. In: Proceedings of SIGMOD 2007, Beijing, June 2007

Xiao Y, Xiong L (2015) Protecting locations with differential privacy under temporal correlations. In: Proceedings of CCS 2015, Denver, October 2015

Giovanni Livraga is an assistant professor at the Computer Science Department of the Università degli Studi di Milano, Italy, where he acquired a PhD in Computer Science. His PhD thesis "Preserving Privacy in Data Release" received the ERCIM STM WG 2015 award for the best PhD thesis on Security and Trust Management in a European University. He has been a visiting researcher at SAP Labs, France, and George Mason University, VA (USA). He has been serving as PC chair and PC member for several international conferences, and as reviewer for several international journals. His research interests are in the area of data protection, privacy, and security, in release, outsourcing, and emerging scenarios. He has collaborated in several national and international projects on different aspects of information protection.

Part II
Microdata Access

Access to European Statistical System Microdata

Aleksandra Bujnowska

1 Introduction

Microdata play an essential role as a primary data source in the production of official statistics. In addition to their use for statistical purposes, the potential of microdata for policy and scientific purposes has been increasingly recognised over recent years. Their analysis being facilitated by technological developments, microdata are extremely valuable as they allow assessment of the underlying structure and causal links of the studied phenomena.

National statistical offices in the European Union (EU) Member States and Eurostat can make microdata available to users for research purposes. While practices to grant access to microdata at national level vary from one country to another, microdata held by Eurostat for all EU Member States (and in some cases European Free Trade Association (EFTA) countries) are provided to researchers according to a transparent approach, in line with applicable legislation.

This chapter focuses on the organisation of access to microdata produced by official statistics, and in particular by the European Statistical System. Sections 2 and 3 explain basic terms and concepts of microdata access. In Sect. 4 the elements of the generic microdata access system are presented. Section 5 then introduces the European microdata access system. Finally, Sect. 6 concludes with some indications on the way forward.

All views expressed are those of the author and not of the Commission (Eurostat).

A. Bujnowska (✉)
European Commission, Eurostat – Statistical Office of the European Union, Luxembourg City, Luxembourg
e-mail: aleksandra.bujnowska@ec.europa.eu

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_6

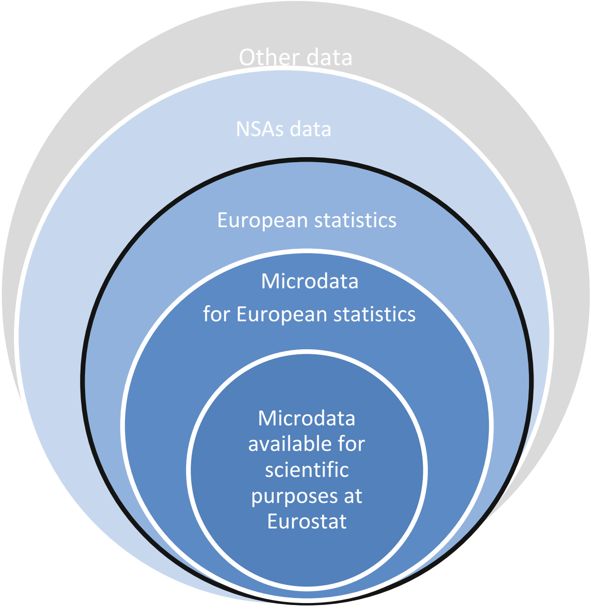

Other data
NSAs data
European statistics
Microdata for European statistics
Microdata available for scientific purposes at Eurostat

Fig. 1 Statistical data available from Eurostat, NSAs and other sources

2 The European Statistical System and European Statistics

The **European Statistical System (ESS)** is a partnership between Eurostat and the national statistical institutes (NSIs) and other national authorities responsible in each Member State for the development, production and dissemination of European statistics. **National statistical authority (NSA) is a generic term for NSIs and other national data providers** (e.g. regional statistical offices, ministries providing administrative data, etc.); a list of NSAs is available on the Eurostat website.[1]

European official statistics are important for EU. They are produced and disseminated by Eurostat in partnership with NSAs. Usually, national official statistics are based on microdata, collected or accessed by NSAs. Microdata are then aggregated, transmitted to Eurostat and published. Where necessary for the production of European statistics, NSAs also transmit microdata to Eurostat (see Fig. 1). Whenever microdata are transmitted, Eurostat may consider granting access to these for scientific purposes. In this way, almost all microdata received by Eurostat are released for scientific purposes.

[1] Path: Eurostat website/About Eurostat/Our partners/European statistical system.

3 Microdata Access Terms and Concepts

Microdata are a form of data where sets of records contain information on individual persons, households or business entities. Traditionally, statistical offices use microdata only to produce aggregated information such as tables. Publication of individual information (microdata) is generally not allowed because it may easily lead to identification of the data subject (person, household or business entity) and therefore to a breach of statistical confidentiality.

Statistical confidentiality is one of the fundamental principles of official statistics. It is the obligation of the statistical offices to protect confidential data.[2] In the context of European statistics, confidential data are data that allow the identification of statistical units (individual persons, households or business entities), thereby disclosing individual information. The statistical unit may be identified in the different forms of statistical output, e.g. the contribution of largest companies may be approximated in business statistics. To prevent this, statistical offices check each output from the point of view of statistical confidentiality. This check is called **statistical disclosure control (SDC).**

The **SDC** methodology helps to identify confidential data in these various output forms and to hide such data, taking into account relationships between the data (e.g. additivity of the tables).

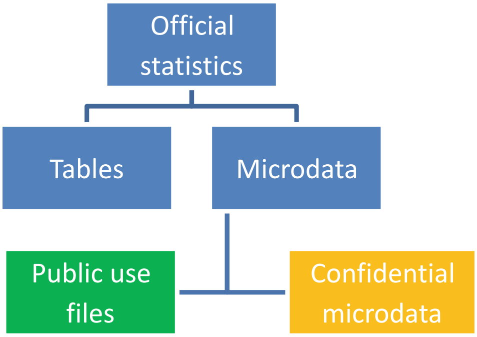
In general, official statistics are available in the form of tables where confidential data are not visible and the data are highly aggregated. But many statistical offices also make available their data in the form of microdata, namely as (see Fig. 2):

- Public-use files accessible to everybody (sometimes upon registration or licence signature)
- Confidential microdata files accessible to researchers satisfying specific access conditions

Confidential microdata files are invaluable for the research community as they allow deep analysis of relationships in the data, i.e. causalities, dependencies, convergences, etc. Microdata access systems were developed by statistical institutes to allow legitimate access to confidential data for scientific purposes.

[2] The ESS and, in a broader sense, official statistics are legally obliged to respect statistical confidentiality. The entities collecting data for purposes other than statistical ones (e.g. commercial, administrative or health purposes) fall into the scope of personal data protection legislation. Statistical confidentiality protection measures are stricter than those stemming from personal data protection legislation.

Fig. 2 Types of data made available by statistical offices



4 Elements of the Generic Microdata Access System

Microdata access systems define under which conditions access to confidential microdata can be granted for external persons, such as researchers. These conditions are normally outlined in legal acts. In the European Statistical System, access to microdata may be granted to researchers carrying out statistical analysis for scientific purposes.[3]

Microdata files may have different levels of detail. The more detailed the data, the easier it is to identify individuals. **Original statistical records** can be easily identifiable as they contain unique direct identifiers such as names, address, social security number or identification number (ID number). These confidential records with direct identifiers are available to the statistical offices only under strict confidentiality protocols.

Microdata without direct identifiers are called **'de-identified' or 'pseudonymised' microdata** (if direct identifiers are replaced by pseudo-identifiers: unique codes replacing all direct identifiers). De-identified microdata with pseudo-identifiers are more and more important for the production of official statistics, as they allow linking data collected from different sources, thus fostering the use of, for example, administrative sources and derivation of further results on the basis of already collected data. Pseudo-identifiers also allow the creation of longitudinal files, following individuals over time. These microdata are still confidential, as the combination of some rare characteristics may lead to identification of unique statistical units.

De-identification is a subprocess of anonymisation. In general, anonymisation is the process of making the data anonymous. However, approaches to this process differ between countries. In some countries, making the data anonymous is defined as removal of names, i.e. de-identification. In the European law, **anonymisation** is defined as the process aiming at complete protection of microdata, such that

[3] Article 23 'Access to confidential data for scientific purposes' of the Regulation (EC) No 223/2209 contains enabling clauses for access to ESS microdata.

the records are no longer identifiable (the records cannot be linked to any 'real' person, household or business entity). The different stages of microdata anonymisation/protection are (see Fig. 3):

- De-identification or pseudoanonymisation: process of removing direct identifiers (such as name, ID number and address) from the confidential data, and replacing them with pseudo-identifiers. Pseudo-identifiers can be used to link datasets.
- Partial anonymisation: application of a set of SDC methods to microdata in order to reduce the risk of identification of the statistical unit. Scientific-use files are the result of partial anonymisation.
- Complete anonymisation: application of SDC methods that completely eliminate the risk of identification of the statistical unit (directly or indirectly). Public-use files contain completely anonymised records.

Table 1 compares all basic types of microdata files and access conditions.

The terms **secure-use files** and **scientific-use files** are specific to the European microdata access system. In the EU countries, there exist similar files but with different names, e.g. scientific-use files are often called 'microdata files for research'. The basic characteristics of these files remain the same:

- Secure-use files are files to which no further methods of statistical disclosure control have been applied. Researchers access these files in the secure environment provided by NSAs (local or remote access). The final results of the work of researchers are checked by NSAs to ensure that they do not reveal confidential data. Each output is checked separately.
- Scientific-use files are files to which methods of statistical disclosure control have been applied to reduce (not to eliminate!) the risk of identification to an appropriate level (partial anonymisation). Researchers have access to such files outside the controlled NSA environment. There are usually no ex post controls by NSAs; researchers need to follow the confidentiality instructions and are responsible for making the published results non-confidential.

Secure use files are the richest form of microdata for research. However, the services related to provision of access are usually expensive for statistical offices. This is because of infrastructure (dedicated environment for on-site or remote access) and operational costs related to output checking.

For statistical offices, scientific-use files seem to be more efficient in terms of cost-benefit ratio. For researchers, the advantage is that they can be used without having to travel to the premises of the statistical offices (or without logging in to a remote, secure system).

Scientific-use files may be standard or tailor made, i.e. adapted to the particular needs of the research project. The risk of a breach confidentiality is smaller if standard files are released than if specific files are produced on request. For researchers, however, the standard files are often not sufficiently detailed (e.g.

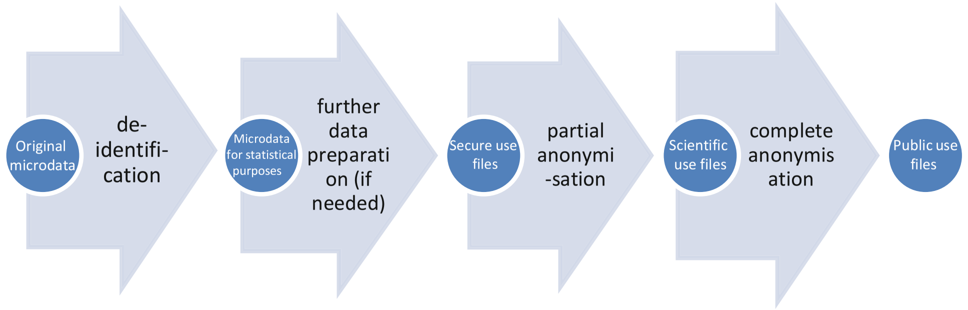


Fig. 3 Anonymisation processes and the resulting types of microdata files

Table 1 Characteristics of the different types of microdata

	Original microdata	Microdata for statistical purposes	Secure use files	Scientific-use files	Public-use files
Identification risk level	Extremely high	Very high	High	Low (reduced)	Eliminated
How the respondents can be identified	By direct identifiers	By combination of indirect identifiers (characteristics such as NUTS[b] level, size class, NACE[c] category)	By combination of indirect identifiers (characteristics such as NUTS[b] level, size class, NACE[c] category)	By combination of indirect characteristics (NUTS[b] level, size class, NACE[c]), but only rare units can be identified	Respondents cannot be identified
Access by	Only limited number of NSA[a] staff	NSA[a] staff	Researchers	Researchers	All
Data availability	Closed NSA[a] production environment	Closed NSA[a] production environment	Dedicated controlled NSA[a] environment for researchers	Data used outside NSA[a] environment, in the premises of the research entities	Data usually available on the NSA[a] website

[a] NSA or Eurostat

[b] NUTS level defines the geographical level of aggregation

[c] NACE categories define the economic activities

the researcher may not need regional details but is interested in the exact age of individuals, whereas the standard files usually provide a medium level of regional details and age in bands).

The scientific-use files released by Eurostat are standard, i.e. they are prepared once for all access requests. Production of tailor-made files would be too burdensome, as the SDC protection measures must be always agreed with the NSAs.

Example of partial anonymisation methods for EU Labour Force Survey (LFS) scientific-use files:

AGE—by 5-year bands
NATIONALITY/COUNTRY OF BIRTH—up to 15 predefined groups
NACE (economic activity)—at 1-digit level
ISCO (occupation)—at 3-digit level
INCOME—provided only as (national) deciles and from 2009
HHNUM—household numbers are randomised per dataset, so that respondents cannot be tracked across time

The most common **SDC methods to anonymise (partially or completely) the microdata files** are:

- Recoding: provision of information at the more general level (e.g. age bands instead of exact age).
- Micro-aggregation: replacement of the original value of the variable (e.g. income) with the average of some (usually 3–5) similar units.
- Record swapping: swapping of, for example, persons between similar households. Swapping adds uncertainty about the identity of the unit in a microdata file.
- Rounding: replacement of original value with rounded figure.
- (Local) suppression: removal of identifying variables in the record or the entire record (e.g. a very large household).
- Sampling: provision of sampled microdata to increase uncertainly about identification as a record referring to particular individual may but does not have to be included in the sample.

The modes of access to secure-use files and scientific-use files are presented in Table 2.

The modes of access listed in Table 2 are complementary and some NSAs provide all options. As the operational costs may be high, the NSA services are sometimes payable.

Table 2 Modes of access to confidential data and respective protection measures

Microdata type used	Mode of access	Confidentiality protection
Secure-use files	– **On site** (separate room, usually in the premises of the NSA where access is provided; researchers can see the data but have no internet access and cannot download or copy the data)	The final results of data analysis are checked for confidentiality (output checking). Each output is checked separately. In some systems, only a sample of output is checked; in others, researchers do it themselves
	– **Remote access** (same functionalities as for on-site access but facilities provided online, no need to travel to the NSA)	
	– **Remote execution** (authorised users submit codes that are executed on the data, but users do not see the data)	The results are checked automatically (‘on the fly’) or manually
	– There might be also combinations of the above modes of access	
Scientific-use files	Files are sent to authorised researchers and used in the premises of research entities	The data are protected (partial anonymisation) before being sent to researchers. Researchers must ensure that the published results do not contain confidential data

5 Use Case: Access to European Statistical System Microdata (European Microdata)

How does the microdata access system work in practice? Eurostat applies a two-step procedure to grant access to microdata for research purposes. In the first step, organisations interested in accessing European microdata submit an application for recognition to Eurostat. In the second step, researchers from recognised research entities submit their concrete research proposals.[4]

Step 1 Recognition as a Research Entity

The recognition of research entities aims at identifying those organisations (or specific departments of the organisations) that carry out research and can be entrusted with confidential data. The assessment criteria refer to the purpose of the entity, its available list of publications and scientific independence. The entities must also describe security measures in place for microdata protection.

The content of the application is evaluated by Eurostat. Upon positive assessment, the head of a recognised research entity signs the commitment that the microdata will be used and protected according to the terms agreed. Eurostat publishes the list of recognised research entities on its website.[5]

To date (2017) more than 700 research entities were recognised. The majority of them are universities and research organisations (see Fig. 4).

Recognition of research entities was introduced by Eurostat to provide a contractual link with the legal entities, rather than with individual researchers.[6]

Step 2 Submission of Research Proposal

In the second step, researchers from recognised entities submit their concrete research proposals to Eurostat. Eurostat then consults all national statistical authorities that provided the data. If an NSA refuses the access, the data of that country are removed from the microdata file.

To be eligible, the research proposal must specify the scientific purpose of the research in sufficient detail, justify the need to use microdata and present the expected outcomes of the research. The results of the research must be made public. Each researcher named in the research proposal as a potential user of the microdata signs an individual confidentiality declaration, in which he or she commits to respect the specific terms of use of confidential data.

In the research proposal, researchers choose the microdata collections they are interested in. In 2017 Eurostat granted access to microdata to 12 data collections (see

[4] The legal basis for access to ESS microdata is Commission Regulation (EU) No. 557/2013 on access to confidential data for scientific purposes. The Regulation defines criteria for eligible research entities and research proposals. It also describes how the microdata shall be made available to researchers (modes of access).

[5] http://ec.europa.eu/eurostat/documents/203647/771732/Recognised-research-entities.pdf.

[6] However, in some national systems, only individual researchers are ‘recognised’.

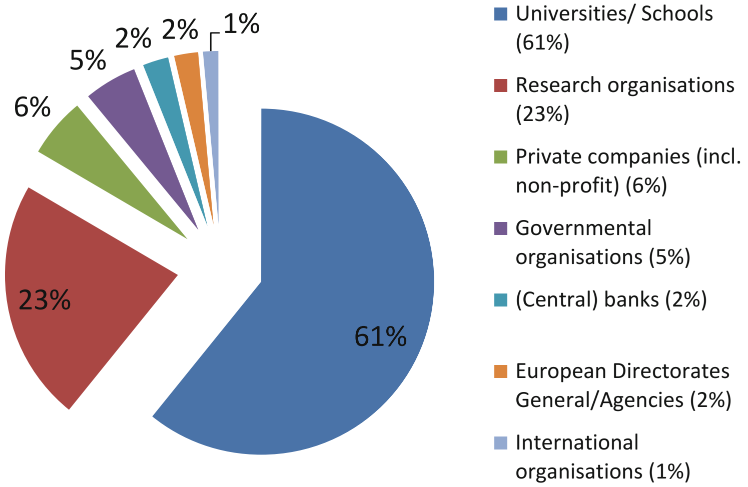


Fig. 4 Types of recognised research entities (in 2017)

Annex 1). Most of the European microdatasets are released as scientific-use files.[7] The datasets most frequently demanded by researchers are EU Statistics on Income and Living Conditions (EU-SILC) and Labour Force Survey (LFS). Together they account for more than 70% of all access requests.

When the research proposal is accepted, the data are made available to the researchers. Researchers may access the data for the period specified in the research proposal. If so requested, researchers receive new releases of the approved microdatasets.

Once the project is finalised, researchers send Eurostat the resulting publications, which are made available on the dedicated website.[8] Researchers must also destroy the confidential data received.

Eurostat receives around 350 applications for access to microdata per year.

6 Conclusions

The ESS microdata access system is specific as it creates a single entry point of access to European microdata owned by the NSAs. NSAs agree on the general access conditions (Regulation 557/2013) and are directly involved in decisions on the release of particular datasets in particular ways (anonymisation method and mode of access), and for particular projects (all NSAs are consulted about each access request).

[7] The anonymisation methods and/or output checking rules are agreed with NSAs.

[8] The publications issued using ESS microdata are available here: https://ec.europa.eu/eurostat/cros/content/publications-received_en.

For Eurostat, access to microdata has become a well-established process. Recently, Eurostat worked on modernising the microdata access system, e.g. launching online forms for microdata access applications and piloting online transmission of scientific-use files. The future plans aim to develop remote execution and to publish more public-use files.[9] Closer collaboration with organisations such as CESSDA (Consortium of European Social Science Data Archives) should contribute to the improvement of microdata access services provided by Eurostat.

Annex 1: European Microdatasets Available for Scientific Purposes

European microdatasets available at Eurostat	Reference years, frequency of new releases	Business (B)/social (S) survey	Microdata file type	Mode of access
1. Adult Education Survey (AES)	2007, 2011	S	Scientific-use file	Off site
2. Community Innovation Survey (CIS)	2002–2012 (bi-annual)	B	Secure-use file and Scientific use file	On site (safe centre in Eurostat) and off site
3. Community Statistics on Information Society (CSIS)	2008–2014 (yearly)	S	Scientific-use file	Off site
4. Continuing Vocational Training Survey (CVTS)	2005, 2010	B	Scientific-use file	Off site
5. European Community Household Panel (ECHP)	1994–2001 (annual)	S	Scientific-use file	Off site
6. European Health Interview Survey (EHIS)	2006–2009 (one data collection depending on the country)	S	Scientific-use file	Off site

(continued)

[9] Currently available European public-use files are published here: https://ec.europa.eu/eurostat/cros/content/puf-public-use-files_en.

European microdatasets available at Eurostat	Reference years, frequency of new releases	Business (B)/social (S) survey	Microdata file type	Mode of access
7. European Road Freight Transport Survey (ERFT)	2011–2014 (annual)	B	Scientific-use file	Off site
8. European Union Statistics on Income and Living Conditions (EU-SILC)	2004–2015 (annual)	S	Scientific-use file	Off site
9. Household Budget Survey (HBS)	2010	S	Scientific-use file	Off site
10. Labour Force Survey (LFS)	1983–2015 (yearly)	S	Scientific-use file	Off site
11. Linked micro-aggregated data on ICT usage, innovation and economic performance in enterprises	2000–2010[a]	B	Secure-use file	On site (safe centre in Eurostat)
12. Structure of Earnings Survey (SES)	1995, 2002, 2006, 2010, 2014	B and S	Secure-use file and scientific-use file	On site (safe centre in Eurostat) and off site

[a] The years covered by the MMD datasets vary from one country to another and are subject mainly to the availability of the Community Innovation Survey and Survey on ICT Usage and e-Commerce in Enterprises data

Aleksandra Bujnowska is a Statistical Officer in Unit B1 ‘Methodology and Corporate Architecture’ at Eurostat. She is leading a team ‘Statistical confidentiality and access to microdata’. For many years, she has been contributing to the development of the European microdata access system and has made several interventions on this subject at various events. She has also coordinated numerous European projects aiming at wider access to confidential data for scientific purposes and at efficient way of micro- and tabular data protection.

Giving the International Scientific Community Access to German Labor Market Data: A Success Story

Dana Müller and Joachim Möller

1 Introduction

In Germany, as in other countries, social security data offer a great opportunity for producing cutting-edge empirical analyses. They are the basis for answering relevant research questions as well as for evaluation studies and evidence-based policy-making. Data resources are especially valuable if they are linked to establishment and individual survey data.

As the research unit of the Federal Employment Agency, the Institute for Employment Research (*Institut für Arbeitsmarkt- und Berufsforschung* (IAB)) is responsible for extracting data from administrative processes to produce micro-datasets that can be used for empirical research on a wide range of labor market topics. In the past, the data were generally kept within the organization, which not only led to a drastic underutilization of the data resources but also limited collaboration projects with national and international academic scholars. There were only rare examples of knowledge spillovers from the international research community to the Institute's research projects. With some major exceptions, researchers at the Institute faced difficulties in keeping pace with the enormous evolution of (micro-) econometric methods. As a consequence, data analysis was mainly descriptive, and publication in refereed international journals was the exception rather than the rule.

D. Müller (✉)
IAB, Nuremberg, Germany
e-mail: dana.mueller@iab.de

J. Möller
IAB, Nuremberg, Germany

University of Regensburg, Regensburg, Germany
e-mail: joachim.moeller@iab.de

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_7

With the growing realization that a closed strategy hinders scientific progress, the strategy was already being softened in the 1990s. In addition, the scientific community showed growing demand for exploiting the valuable data resources to answer research questions. This outside pressure favored the process of opening; however, development took some time. There was no standardized and institutionalized way for researchers outside the IAB to gain access to the data until 2004 (Kohlmann 2005).

There were two important impulses to improve data access for the scientific community. The first was the labor market reforms implemented between 2003 and 2005. An element of these reforms was to strengthen scientific evaluation of active labor market instruments to increase their efficiency. The second was the recommendation of the German Commission on Improving the Information Infrastructure Between Science and Statistics (*Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik*) to establish a research data center at each public producer of microdata in Germany. The Federal Employment Agency (*Bundesagentur für Arbeit* (BA)) followed this recommendation and established a research data center within the IAB in spring 2004. This was facilitated by the Federal Ministry of Education and Research, which funded the initial process for 3 years. After an evaluation by the German Data Forum (*Rat für Sozial- und Wirtschaftsdaten*) in 2006, a research data center was established as a permanent department of the IAB (see Solga and Wagner 2007). Today, the Research Data Centre of the Federal Employment Agency at the IAB (RDC-IAB) is one of 31 such research data centers that have been established (see http://www.ratswd.de/en/data-infrastructure/rdc).

The establishment of the RDC-IAB and the reorganization of the institute under the directorship of Jutta Allmendinger were the starting signal for numerous collaborations with external scholars. IAB researchers profited especially from joint projects with international partners. Some of these projects led to publications in top-ranked international journals (see Dustmann et al. 2009; Card et al. 2013, among others). Active labor market policies were evaluated using the latest empirical methods (see, for instance, Wolff and Stephan 2013). In several cases, the labor market research based on the RDC-IAB data led to a new design of active labor market policy. A recent example is an evaluation study on the compulsory integration agreement between the jobseeker and the caseworker. Using a randomized field experiment and following the labor market biographies of the persons included in the experiment, IAB was able to show that for some groups of unemployed, the compulsory regulation is counterproductive and should be replaced by a more flexible handling of the instrument (van den Berg et al. 2016).

Another important field is to monitor and evaluate the effects of new labor market regulations or institutions. One example is the minimum wage that was first implemented in the German construction industry in 1997 and later extended to other sectors (König and Möller 2009; Möller 2012; Aretz et al. 2013). The various effects of the general statutory minimum wage that was implemented on 1 January 2015 are currently being analyzed in several projects based on RDC-IAB data. In general, labor administration and policy-makers have been profiting from better insight into labor market structures and are now able to optimize labor market processes and instruments.

Through data access points that adhere to the highest standards of data security and confidence, the RDC-IAB provides researchers with access to its data resources not only in Germany but also in the United Kingdom (UK) and the United States (US). The number of users is steadily increasing. In 2016, almost one-third of all data use agreements were from a non-German facility. In the future, the RDC-IAB will expand the possibilities of data access even further. Mutual access to microdata of different European countries or linkage of different datasets also requires new technical solutions.

The aim of this chapter is to provide an overview over the activities of the RDC-IAB and developments planned for the future. It begins with a description of the core data and the modes of data access. It then describes how demand for the data evolved over time. Further infrastructural developments and research activities are described in Sects. 5 and 6. Section 7 concludes.

2 The Research Data Centre at the IAB and Its Data Resources

The RDC-IAB is primarily a service-oriented department but also conducts its own research projects and acquires grants from various foundations. It provides access to high-quality microdata for researchers in Germany and abroad, in compliance with German data protection legislation. To accomplish this aim, the RDC-IAB performs the following tasks:

- It develops specific data products with high research potential for labor market studies; this includes the necessary preparation and harmonization of the raw data as well as regular updates.
- It compiles detailed documentation of the data products and considers technical aspects of the data as well as statistical properties. It provides tools to facilitate analyses of microdata and offers individual counseling.
- To prevent re-identification of personal information, the RDC-IAB develops and applies anonymization strategies.
- It develops standardized ways for (inter)national researchers to access data.
- It promotes data products and data access through active participation in (inter)national workshops, conferences, and seminars.
- It conducts its own research on and with the available data products to improve their quality and to assess their research potential and ability to provide competent individual counseling for external researchers.

The RDC-IAB shares its activities on data access and the development of new data products, metadata, and research with an international network of research data centers, data providers, and scientific institutions.

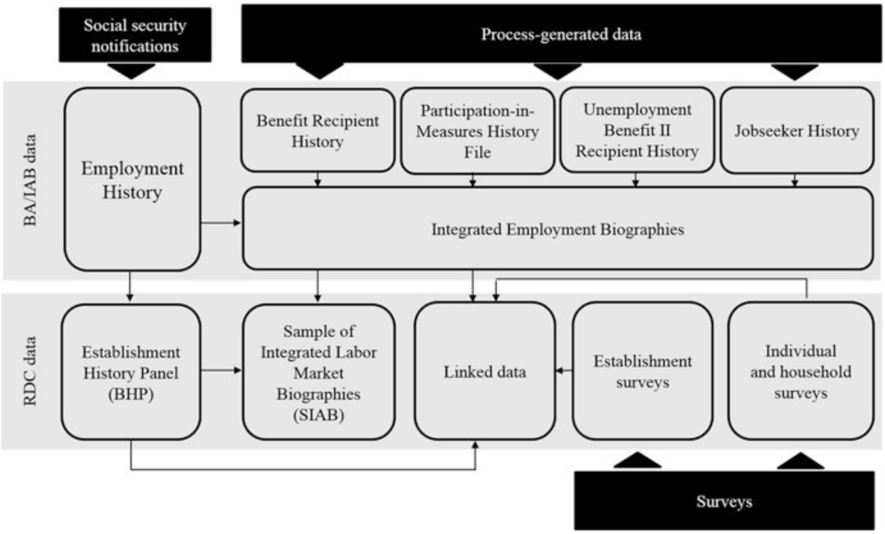

Social security notifications
Process-generated data
BA/IAB data
Employment History
Benefit Recipient History
Participation-in-Measures History File
Unemployment Benefit II Recipient History
Jobseeker History
Integrated Employment Biographies
RDC data
Establishment History Panel (BHP)
Sample of Integrated Labor Market Biographies (SIAB)
Linked data
Establishment surveys
Individual and household surveys
Surveys

Fig. 1 BA/IAB data sources and core data products

Thirteen years after its foundation, the RDC-IAB is considered the most important supplier of German labor market microdata. Currently, 16 data products on individuals, households, and establishments are available to the scientific community. The data originate from administrative data from the social security system's notification process and internal processes of the BA and from surveys conducted by the IAB. The RDC-IAB enlarges the research potential by linking existing data with other administrative data or surveys. In 2011, the Record Linkage Center was founded at the IAB, a joint project with the University of Duisburg-Essen that was funded by the German Research Foundation (*Deutsche Forschungsgemeinschaft*). The methods developed in this context facilitate the linkages of microdatasets without a unique identifier.

The RDC-IAB offers labor market microdata on individuals, households, and establishments. These datasets are generated from three different sources: (i) register data from the social security system's notification process, (ii) data from internal procedures of the BA, and (iii) survey data (see Fig. 1 for an overview).

The legal basis for social security data collection is provided by the German Data and Transmission Act (*Verordnung über die Erfassung und Übermittlung von Daten für die Träger der Sozialversicherung*) and the Social Act (Social Code Book IV). As part of the social security notification procedure, all employers are required to report several items and characteristics of their employees. In principle, two kinds of information are stored. The first is information that is collected for statistical purposes, and the second is information that is collected to compute the amount of social security contributions and the resulting claims. The administrative data from the internal procedures of the BA are the result of the agency's fulfillment of tasks in accordance with the Social Code Books II and III. These are the administration of the compulsory unemployment insurance, calculations of unemployment benefits and the corresponding entitlement periods, consultation sessions with the unemployed, placement offers, and active labor market measures. The collection of these administrative data began in 1975. The IAB generates historical data from these records and combines them into a comprehensive unique dataset, i.e., the Integrated Employment Biographies (IEB). Not all variables are available for the entire observation period. Due to changes in statutory regulations, administrative data sources start at different points in time. The RDC-IAB updates its data products regularly and offers different samples of these rich administrative data sources for research purposes.

As well as processing the administrative microdata, IAB conducts various surveys. In addition, the RDC-IAB exploits the opportunity to link these surveys to administrative data. According to German data protection rules, this is allowed if the respondents consent to the linkage (Heining 2010).

Currently, the RDC-IAB provides 16 datasets (see Table 1). A short description of selected examples for establishment, individual, and household data is given in the paragraphs below. More detailed information is available on the website of the RDC-IAB (see http://fdz.iab.de/en.aspx).

The IAB Establishment Panel (IABB) is an annual representative survey of approximately 16,000 establishments in Germany (Bellmann 2002; Fischer et al. 2009; Ellguth et al. 2014). The survey started in 1993 for West Germany and has covered East Germany since 1996. It includes only establishments with at least one employee covered by social security on 30 June of the previous year. The IABB contains various topics, such as the development of total employment, business policy, investments, export, innovations, personnel structure, apprenticeship and vocational training, recruitments and dismissals, wages, working hours, training programs, and alternating annual topics.

The IAB Establishment History Panel (BHP) is a yearly cross-sectional dataset on all establishments in Germany with at least one employee eligible for social security contributions (Spengler 2008; Eberle and Schmucker 2017). The dataset is a 50% random sample drawn from establishment identification numbers and gives information for the reference date (30 June) of each year. The panel starts in 1975 for West Germany and in 1992 for East Germany and includes between 640,000 and 1.5 million establishments per year. The BHP contains information on workforce composition such as gender, age, nationality, occupational status and qualification as well as branch of industry and the location of the establishment. Furthermore, there is information on worker in- and out-flows and indicators of establishment entries and exits (Hethey-Maier and Schmieder 2013).

The German Management and Organizational Practices Survey (GMOPS) is a novel establishment dataset provided at the RDC-IAB since September 2016 (Broszeit and Laible 2016). GMOPS, funded by the Leibniz Association, belongs to Management Practices, Organizational Behavior, and Firm Performance in Germany, a collaboration project that was jointly carried out by the IAB, the Kiel Institute for the World Economy (IfW), and the Institute for Applied Social Sciences (infas). The survey is based on the US Census Bureau's "Management

Table 1 Data products of the Research Data Centre at the IAB

Dataset	Administrative Data	Survey	Linked Data
Establishment data			
IAB Establishment Panel		✓	
IAB Establishment History Panel (BHP)	✓		
German Job Vacancy Survey of the IAB		✓	
German Management and Organizational Practices (GMOP) Survey		✓	
Individual data/household data			
Sample of Integrated Labor Market Biographies (SIAB)	✓		
Panel Study‘ Labor Market and Social Security (PASS)		✓	✓
Working and Learning in a Changing World		✓	
Biographical Data of Social Insurance Agencies in Germany (BASiD)	✓		✓
IAB-SOEP Migration Sample (IAB-SOEP MIG)		✓	✓
Employee survey‘ bonus payments, wages increases, and fairness (BLoG)		✓	
IZA/IAB Administrative Evaluation Dataset	✓		✓
lidA—*leben in der Arbeit*. German cohort study on work, age, and health		✓	
S-MGA—The Study on Mental Health at Work		✓	
Integrated establishment and individual data			
Linked Employer–Employee Data from the IAB (LIAB)		✓	✓
Linked Personnel Panel (LPP)		✓	✓
Panel “WeLL”—Employee Survey for the Project “Further Training as a Part of Lifelong Learning”		✓	✓

and Organizational Practices Survey” (MOPS) from 2010. Large parts of the questionnaire were translated into German, and additional information has been added, for example, on work–family balance and health promotion and on sales, export, and innovation. The survey was conducted once, in the period 2014–2015. The information in the data relates to the years 2008 and 2013 and covers 1927 establishments.

The “Sample of Integrated Labor Market Biographies” (SIAB) is a 2% random sample from the Integrated Employment Biographies (Dorner et al. 2010). The employment biographies cover the period from 1975 until 2014 for West Germany and from 1992 until 2014 for East Germany. The microdata include more than 1.7 million individuals in total and cover day-exact information on sociodemographic characteristics, employment, benefit receipts and job searches, and location and establishment.

The "Panel Study Labor Market and Social Security" (PASS) is an annual household survey in the field of labor market, welfare state, and poverty research in Germany (Trappmann et al. 2013). The survey consists of two random samples. The first sample includes households and individuals receiving means-tested social assistance (the so-called Unemployment Benefit II), and the second includes any other households of German residents. The field phase of the first wave ran from December 2006 to July 2007. Both random samples are continued over time. To guarantee representativeness in each wave for Unemployment Benefit II recipients, refreshment samples of households that claimed Unemployment Benefit II for the first time were drawn for the following waves. The survey includes a personal interview with the head of household and, subsequently, personal interviews with all members of the household aged 15 or older. Persons aged 65 or older are interviewed with a reduced questionnaire. The last wave of 2015 includes approximately 13,300 persons in nearly 9000 households. More than 11,700 of these persons and more than 7800 of these households have been interviewed multiple times.

The "Linked Employer–Employee Data" from the IAB (LIAB) combines the IAB Establishment Panel with data for employees from the Integrated Employment Biographies (Heining et al. 2014). The LIAB is useful for the simultaneous analysis of supply and demand on the German labor market. There are two different versions of the LIAB. The LIAB cross-sectional model contains all waves of the IAB Establishment Panel and linked information of all employees on 30 June of a given year. The updated LIAB longitudinal model is a sample of establishments repeatedly interviewed between 2000 and 2011 and is linked to all employees who worked at least 1 day in one of the establishments included. The employment biographies begin in 1992 and continue until 2014. Additional generated variables comprise the employment and unemployment experience before 1992.

The "Linked Personnel Panel" (LPP) is another novel-linked employer–employee dataset on human resource work, corporate culture, and management instruments in German establishments (Bellmann et al. 2015). It evolved from the "Quality of Work and Economic Success" project, a collaboration between the IAB, the University of Cologne, and the Centre for European Economic Research (ZEW). It is funded by the IAB and the Federal Ministry of Labor and Social Affairs (*Bundesministerium für Arbeit und Soziales* (BMAS)). The project is designed to include three survey waves of employers and their employees, at 2-year intervals. The current data product contains the first two waves. In the first wave (2012/2013), 1219 establishments and more than 7500 of their employees were interviewed. The second wave (2014/2015) contains information for 771 establishments and approximately 7280 employees. The LPP Employer Survey is directly attached to the IAB Establishment Panel; therefore, all information of the IAB Establishment Panel can be included.

For each data product, the RDC-IAB provides detailed documentation in German and English. There are two publication series. The *FDZ Datenreport* series contains documentation of the data, changes to previous versions and information on data preparation, as well as methodological aspects on data handling. Additional information on frequencies, labels, or working tools is available on the website of the RDC. Currently, the RDC-IAB is working on transferring from PDF format documentation to a web application for all data documentation using the DDI standard.[1] The *FDZ Methodenreport* series addresses methodological aspects and problems. It may be used as a publication outlet by any author working with BA or IAB data.

3 Data Access

The legal basis for data access is found in §75 of the German Social Code Book X and §282 (7) of the German Social Code Book III. The need for data protection determines the ways in which data can be accessed. In general, this means that the more detailed the data are, the more restricted the access to the data is. The RDC-IAB offers four kinds of data access for the scientific community:

1. *Campus files* are fully anonymized and useful only for teaching. Users need to register and agree to terms of use before the campus file can be downloaded.
2. *Scientific use files* are de facto anonymized microdata that are submitted to scientific institutions in Germany and EU Member States within the scope of §282(7) of the German Social Act III. The information has been reduced for data confidentiality reasons to the extent that re-identification of personal information would be possible only with a disproportionate amount of time, expense, and effort (Hochfellner et al. 2014). Scientific use files are offered to researchers for research projects in the field of labor market research but not for teaching or for commercial research interests. Data security must be guaranteed by the scientific institution applying for the data.
3. The anonymization of the scientific use file restricts the research potential; therefore, the RDC-IAB offers *weakly anonymized data* (i.e., de-identified microdata) with more detailed information. Access is possible only via *on-site use* within the scope of §75 of Social Code Book X. The RDC-IAB provides separate workplaces within a secure computing environment in Nuremberg and at various locations in Germany, the USA, and the UK (Bender and Heining 2011). Within the secure computing environment, researchers have direct access to weakly anonymized data; however, they can obtain the output of their programs only after disclosure reviews by RDC staff (for details, see Hochfellner et al. 2014). On-site use is limited to research projects in the field of social benefits or labor market research.
4. *Remote execution* means that researchers prepare their programs with artificial data and upload the programs in the Job Submission Application (JoSuA), which is described in more detail in Sect. 5. Researchers never view the original data. They receive their results after a disclosure review by RDC staff. Remote execution is also possible after on-site use of the data.

[1] DDI stands for the Data Documentation Initiative. See https://www.ddialliance.org/.

The use of scientific use files, remote execution, and on-site access must fulfill certain requirements in accordance with the legal regulations (for more details, see Hochfellner et al. 2014). Therefore, the RDC-IAB offers standardized request forms for all data access to clarify whether or not the research purpose complies with the legal requirements. Final permission for on-site use is granted by the BMAS. After a request has been approved, a contract of data use for a specific project within a specific period is concluded between the researcher's institution and the RDC-IAB. The contract specifies the data protection rules and severe sanctions in the event that these rules are violated.

Note that some of the datasets listed in Table 1 are available only for on-site use. Among others, this applies for all linked datasets.

4 Development of the Demand for Data Products

The RDC's data products enjoy immense popularity in Germany and abroad. Figure 2 shows the number of users and the numbers of projects for each year since 2005. Generally, more than one researcher works on one project, and the duration of a project usually exceeds 1 year. The number of users has increased consistently over time. In 2015, for instance, the RDC-IAB reached just over 1000 users, who work, or were working, on 514 projects. In 2016, the number of users and projects was higher still.

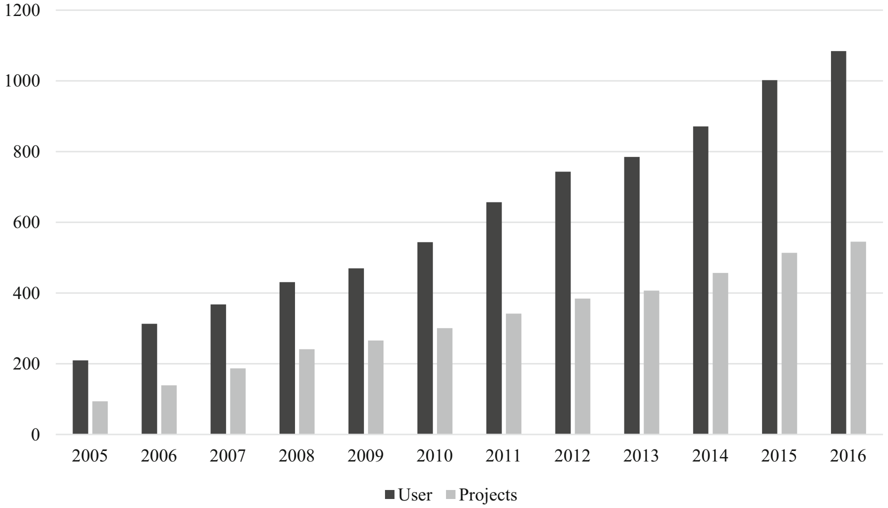


Fig. 2 Development of the number of data product users and number of projects, RDC-IAB, 2005–2016

Most of the users of RDC-IAB data products work at a German research institute or university, but a growing number of data requests are from international scholars. The noticeable increase in the number of users was made possible through the project "The Research-Data-Centre in Research-Data-Centre Approach: A First Step Towards Decentralised International Data Sharing" (RDC-in-RDC),[2] funded from 2011 to 2013 by the Ministry of Education and Research (*Bundesministerium für Bildung und Forschung* (BMBF)), follow-up funded by the National Science Foundation under program SES-1326365, and financially supported by the project "Data Without Boundaries" within the Seventh Framework Programme of the European Union (Heining and Bender 2012). Before this project began, the only location with access to weakly anonymized data was the RDC-IAB in Nuremberg. Capacity was limited to five workstations. Within the project, data access was established at various locations in research data centers or institutions that offer a data protection infrastructure similar to that of the RDC-IAB. In each partner organization, there is a secure guest room, and researchers are provided data access via a secure internet connection to the RDC-IAB in Nuremberg. In principle, there are no differences between the international and German RDC-in-RDC IAB approaches in either technical implementation or the application process. However, the difference in the legal framework must be considered. This legal framework requires that only (de facto) anonymized data can be accessed from abroad. Therefore, RDC staff construct (de facto) anonymized datasets for approved projects. Figure 3 shows,

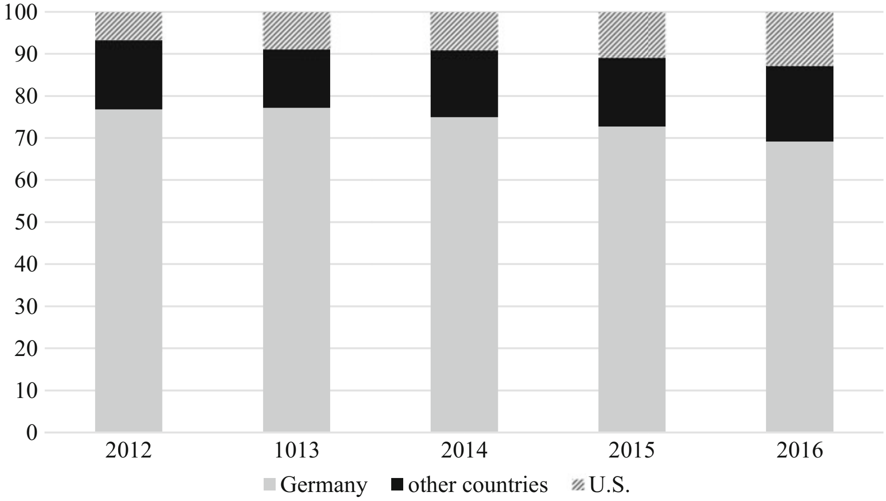


Fig. 3 Contractual partners of the IAB Research Data Center by country, 2012–2016

[2] This project was initiated by Stefan Bender, former head of the RDC.

Table 2 Number of projects by datasets and year, 2012–2016

Dataset	2012	2013	2014	2015	2016
SIAB	69	107	154	203	231
LIAB	77	85	101	111	132
IABB	90	74	71	76	75
PASS	34	39	47	49	54
BHP	28	30	37	41	47
ALWA	26	29	28	24	23
BASiD	7	11	14	21	23
All others	100	81	71	79	78
Total	431	456	523	604	663

Note: For explanations of dataset acronyms, see text

for example, that 31% of all research projects in 2016 were from a non-German facility.

The demand for several RDC-IAB data products depends on the user's research purposes. Table 2 shows a list of the seven most commonly requested datasets within the past 5 years. The number of projects in Table 2 differs from the number of projects in Fig. 1 because the number of data products within a project is not restricted to one data product. The SIAB has been the most requested dataset since 2013. The SIAB is available both as a scientific use file and for on-site use. It includes individual information covering the period since 1975 and is enriched by several characteristics of the employer. The SIAB, therefore, suits a wide range of research purposes in the fields of labor market research and social research. It comes as no surprise that the Linked Employer–Employee dataset LIAB ranks second. Here, valuable and reliable information from the administrative data is combined with comprehensive information from the establishment panel. These rich and innovative data hold enormous research potential.

The figures and tables above show the high demand for RDC-IAB data products. The importance for the scientific community is also demonstrated by publication records. The RDC-IAB literature database not only includes dataset descriptions and methodology reports but also lists the publications by researchers using the data offered by the RDC-IAB (see http://fdz.iab.de/en/FDZ_Publications/FDZ_Literature_Database.aspx). By 2015, the list contained 85 publications in journals, most of which (68) are in renowned refereed journals; there were also 90 monographs, 61 working papers/discussion papers, and 13 articles in collected editions (previous statistics are available in Bender et al. 2011). Note that it is likely that the number of publications is underreported because, unfortunately, not all users comply with the obligation to inform the RDC-IAB about their publications if RDC-IAB data sources are used.

5 Innovative Infrastructure for Data Access

The RDC-IAB works on its infrastructure continuously to improve the ways in which data can be accessed and processed. One of the most important infrastructure projects was the recent implementation of a remote data submission environment at the RDC. Most user projects work with weakly anonymized data, access to which is restricted to on-site use and the remote execution of data processing programs. The growing number of users increased the number of jobs submitted for remote executions and disclosure review for on-site use. In 2014, for instance, the RDC-IAB reached a total of approximately 1800 remote jobs. As a result, the RDC-IAB reached its own capacity limits more frequently. Therefore, it was necessary to improve the infrastructure and thereby manage remote data execution and the disclosure review service in a more automated way. The RDC-IAB decided to implement the *Job Submission Application* (JoSuA) environment in 2015 (Eberle et al. 2017). The software is maintained by the Institute of Labor Economics (IZA). The main innovation designed for JoSuA is to provide separate modes of job submission. The motivation is that typical research work requires a lot of data processing and testing for project-internal purposes only. Consequently, the bulk of output of these procedures is not suitable for use in a publication or presentation, which means that it is not necessary to export most of the output. Therefore, JoSuA provides two kinds of job submission. The first mode, "Internal Use," is completely automated and should be used to prepare the data and test the empirical methods. In this mode, the manual output controls are replaced by a script-based automated disclosure review. The initial text output files are converted into image files and can only be previewed in JoSuA; downloading is not possible. The second mode, "Presentation/Publication," should be used after data preparation and testing are finished. In this mode, do-files and output files are manually reviewed for disclosure risk by scientific staff members of the RDC-IAB and are subsequently made available for download via JoSuA.

Once a new job is submitted in "Publication/Presentation" mode, the previous "Internal Use" jobs are inaccessible.

The main advantages of these distinct job submission modes are obvious:

- The environment increases data security and minimizes the risk of disclosure.
- The results in "Internal Use" can be directly checked by the researcher without significant delay.
- Only output required for a presentation or publication is given to the researcher.
- All other output remains within the secure computing environment of the RDC.

Figure 4 shows the number of jobs for both modes since JoSuA was made available to RDC-IAB users. On the one hand, the number of jobs has increased enormously. In 2014, the number of jobs per month was less than 170. This number has since almost quadrupled in a typical month. On the other hand, due to JoSuA, it was possible to limit the volume of output to be checked for any disclosure risks by a scientific staff member of the RDC-IAB, since the jobs in

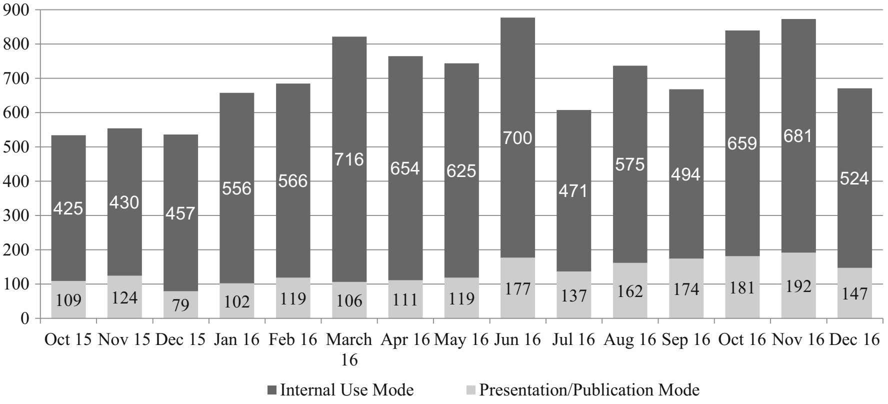


Fig. 4 Number of jobs via JoSuA at RDC-IAB by execution mode, October 2015 to December 2016

"Presentation/Publication" mode were only a fraction of the total. Hence, with equal personnel resources, a more intensive use of the datasets was made possible. However, due to the increase in general demand, the capacity limits of the RDC-IAB have again been reached.

A further project aimed to improve searching on the RDC-IAB datasets. To this end, a metadata management and web information system using the DDI standard were implemented. Both applications are currently loaded with the relevant contents.

Furthermore, future infrastructure development includes an enhancement and extension of the RDC-in-RDC IAB approach. The RDC-IAB plans further on-site locations both within and outside Europe. Enhancement of the concept would foster the mutual exchange of microdata access possibilities with partner institutions. Concurrently, the RDC-IAB will be involved in two feasibility studies to extend opportunities for data access via remote access. The first is planned in the framework of an extension of the Virtual Research Environment project, which was developed to support collaborative use of microdata in joint projects of spatially distributed research institutions. The Virtual Research Environment was used by the joint project "Reporting on socioeconomic development in Germany—soeb 3" (*Forschungsverbund Sozioökonomische Berichterstattung* 2017). The second feasibility study is planned jointly with the Center for Urban Science and Progress (CUSP) at New York University.

Finally, the RDC-IAB plans to elaborate on concepts for creating data access for linked microdatasets that cannot be stored at one of the data providers involved because of data protection, ownership claims, or other restrictions. One technical solution could involve linking the datasets held by two or more data providers and analyzing them on an encapsulated high-security server of a data custodian "on the fly" (or in the cloud) via remote access. The linkage of the datasets according to such a concept is only temporary in the workspace of the server and remains in place only as long as needed for the statistical analyses to be completed.

6 Research Activities

A general principle of the RDC-IAB is that its staff members should not be engaged exclusively in data provision services but that they should also conduct their own research, at least to a limited extent. The idea behind this is that working on new data products and data quality requires research experience. Finally, research experience is helpful for better individual data counseling.

Many of the research activities of the RDC's staff members are based on collaborations with national and international external researchers (e.g. Card et al. 2013; Hirsch et al. 2014; Bender et al. 2016; Fackler et al. 2016). This guarantees an intensive exchange of knowledge on new econometric and statistical methods, new trends in data handling, or the improvement of data collection and survey techniques. In addition, the RDC-IAB requires that its research activities be presented at national and international workshops and conferences. Furthermore, the RDC-IAB is involved in numerous external projects that are funded by the Germany Research Foundation, the Federal Ministry of Education and Research, or the Ministry of Labor and Social Affairs, for instance. These projects are frequently carried out in collaborations with universities, other research institutes, or research data centers. All projects are described in detail on the website of the RDC-IAB.[3]

7 Conclusions

This paper describes the opening of German administrative labor market microdata to the international scientific community as a great success story. In 2004, the Research Data Centre at the IAB was established to offer rich administrative microdata samples of integrated labor market biographies, linked employer–employee information, and other data to a network of international scholars. All sides have profited from abandoning the closed strategy for social security data that prevailed in the past. The German Federal Employment Agency, the Ministry of Labor and Social Affairs, and other stakeholders of the IAB have benefited from improved evidence-based policy advice, the international scientific community from new opportunities to answer relevant labor market research questions with reliable and comprehensive data, and, last but not least, the IAB itself through a number of joint projects and collaboration with international researchers. An important point is that access to the data complies with strict German data protection rules. To this end, the

[3] See http://fdz.iab.de/en/FDZ_Projects/projects.aspx/Bereichnummer/17.

Research Data Centre at the IAB has not only improved anonymization techniques but also established a data access infrastructure that meets the demanding requirements. An example is an environment that allows remote data processing. Future developments include mutual microdata exchange between partner institutions and improvements in data linkage techniques in conformity with data protection rules.

Acknowledgements For helpful comments and discussions, the authors would like to thank Daniela Hochfellner and Alexandra Schmucker.

References

Aretz B, Arntz M, Gregory T (2013) The minimum wage affects them all: evidence on employment spillovers in the roofing sector. Ger Econ Rev 14(3):282–315. https://doi.org/10.1111/geer.12012

Bellmann L (2002) Das IAB-Betriebspanel: Konzeption und Anwendungsbereiche. Allg Stat Arch 86(2):177–188

Bellmann L, Bender S, Bossler M et al (2015) LPP – Linked Personnel Panel – quality of work and economic success: longitudinal study in German establishments (data collection on the first wave). FDZ-Methodenreport 05/2015. http://doku.iab.de/fdz/reporte/2015/MR_05-15_EN.pdf. Accessed 14 Sept 2017

Bender S, Heining (2011) The Research-Data-Centre in Research-Data-Centre approach: a first step towards decentralised international data sharing. FDZ-Methodenreport 07/2011. http://doku.iab.de/fdz/reporte/2011/MR_07-11_EN.pdf. Accessed 14 Sept 2017

Bender S, Dieterich I, Hartmann B et al (2011) FDZ-Jahresbericht 2009/2010. FDZ-Methodenreport 06/2011. http://doku.iab.de/fdz/reporte/2011/MR_06-11.pdf. Accessed 14 Sept 2017

Bender S, Bloom N, Card D et al (2016) Management practices, workforce selection and productivity. CEP Discussion Paper No 1416. Centre for Economic Performance, London

Broszeit S, Laible M-C (2016) German Management Organizational Practices survey (GMOP 0813): data collection. FDZ-Methodenreport 06/2016. http://doku.iab.de/fdz/reporte/2016/MR_06-16_EN.pdf. Accessed 14 Sept 2017

Card D, Heining J, Kline P (2013) Workplace heterogeneity and the rise of West German wage inequality. Q J Econ 128(3):967–1015

Dorner M, Heining J, Jacobebbinghaus P et al (2010) The sample of integrated labour market biographies. Schmollers Jahrbuch 130(4):599–608

Dustmann C, Ludsteck J, Schönberg U (2009) Revisiting the German wage structure. Q J Econ 124(2):843–881

Eberle J, Schmucker A (2017) The establishment history panel: redesign and update 2016. Jahrb Natl Okon Stat. https://doi.org/10.1515/jbnst-2016-1001

Eberle J, Müller D, Heining J (2017) A modern job submission application to access IABs confidential administrative and survey research data. FDZ-Methodenreport 01/2017. http://doku.iab.de/fdz/reporte/2017/MR_01-17_EN.pdf. Accessed 14 Sept 2017

Ellguth P, Kohaut S, Möller I (2014) The IAB establishment panel: methodological essentials and data quality. J Labour Market Res 47(1/2):27–41

Fackler D, Schnabel C, Schmucker A (2016) Spinoffs in Germany: characteristics, survival, and the role of their parents. Small Bus Econ 46(1):93–114

Fischer G, Janik F, Müller D et al (2009) The IAB establishment panel: things users should know. Schmollers Jahrbuch 129(1):133–148

Forschungsverbund Sozioökonomische Berichterstattung (2017) Berichterstattung zur sozioökonomischen Entwicklung in Deutschland: Exklusive Teilhabe – ungenutzte Chance. W. Bertelsmann, Bielefeld

Heining J (2010) The Research Data Centre of the German Federal Employment Agency: data supply and demand between 2004 and 2009. J Labour Market Res 42(4):337–350

Heining J, Bender S (2012) Technical and organisational measures for remote access to the micro data of the Research Data Centre of the Federal Employment Agency. FDZ-Methodenreport 08/2012. http://doku.iab.de/fdz/reporte/2012/MR_08-12_EN.pdf. Accessed 14 Sept 2017

Heining J, Klosterhuber W, Seth S (2014) An overview on the linked employer–employee data of the Institute for Employment Research (IAB). Schmollers Jahrbuch 134(1):141–148

Hethey-Maier T, Schmieder JF (2013) Does the use of worker flows improve the analysis of establishment turnover? Evidence from German administrative data. Schmollers Jahrbuch 133(4):477–510

Hirsch B, Jahn EJ, Toomet O et al (2014) Does better pre-migration performance accelerate immigrants' wage assimilation? Labour Econ 30:212–222

Hochfellner D, Müller D, Schmucker A (2014) Privacy in confidential administrative micro data: implementing statistical disclosure control in a secure computing environment. J Empir Res Hum Res Ethics 9(5):8–15

Kohlmann A (2005) The Research Data Centre of the Federal Employment Service in the Institute for Employment Research. Schmollers Jahrbuch 125(3):437–447

König M, Möller J (2009) Impacts of minimum wages: a microdata analysis for the German construction sector. In: Blien U, Jahn E, Stephan G (eds) Unemployment and labour market policies: novel approaches. Emerald, Bingley, UK, pp 716–741

Möller J (2012) Minimum wages in German industries: what does the evidence tell us so far? J Labour Market Res 45(3/4):187–199

Solga H, Wagner GG (2007) A modern statistical infrastructure for excellent research and policy advice: report on the German Council for Social and Economic Data during its first period in office (2004–2006). Working Paper Series. German Council for Social and Economic Data, Berlin

Spengler A (2008) The establishment history panel. Schmollers Jahrbuch 128(3):501–509

Trappmann M, Beste J, Bethmann A et al (2013) The PASS panel survey after six waves. J Labour Market Res 46(4):275–281

Van den Berg GJ, Hofmann B, Stephan G, Uhlendorff A (2016) Eingliederungsvereinbarungen in der Arbeitslosenversicherung: nur ein Teil der Arbeitslosen profitiert von frühen Abschlüssen. IAB-Kurzbericht 03/2016. http://doku.iab.de/kurzber/2016/kb0316.pdf. Accessed 14 Sept 2017

Wolff J, Stephan G (2013) Subsidized work before and after the German Hartz reforms: design of major schemes, evaluation results and lessons learnt. IZA J Labor Policy 2:1–24. https://doi.org/10.1186/2193-9004-2-16

Dana Müller is head of the Research Data Center (FDZ) of the Federal Employment Agency at the Institute for Employment Research since October 2016. She studied Sociology at the Chemnitz University of Technology and worked as a researcher at the FDZ.

Joachim Möller Studies of Philosophy and Economics at the Universities of Tübingen, Strasbourg and Konstanz. Doctorate: University of Konstanz, 1981 and Habilitation 1990. Current Position: Full Professor of Economics at the University of Regensburg, Director of the Institute for Employment Research of the Federal Employment Agency (IAB).

Hungary: A Case Study on Improving Access to Administrative Data in a Low-Trust Environment

Ágota Scharle

1 Introduction

Post-socialist European Union (EU) member states share a strong tradition of extensive data collection by the state, but typically, administrative data are not systematically used to support evidence-based policymaking. In most cases this is mainly due to the relatively low efficiency of governance (implying that governments do not generate much demand for evaluation) and in some cases also to overly strict legislation on personal data protection. Access to microdata may be further constrained by lack of trust between academic and government organisations, as well as within the government.

The Hungarian case is a good example of improving access to administrative data for research and policy analysis in such a context. This chapter focuses in particular on three issues: (1) the interests of stakeholders involved in a legislative process that yielded a new law on microdata access in 2007, (2) the negotiation process leading to the new law and (3) how particular features of the new law satisfied opponents while meeting the demands of data users. To illustrate the impact and sustainability of the new legislation, the paper also briefly describes the outcomes in terms of summary statistics on data requests since 2007 and some examples of how administrative data have been used by researchers and policymakers to inform policy discussions since 2007.

Á. Scharle (✉)
Budapest Institute for Policy Analysis, Budapest, Hungary
e-mail: agota.scharle@budapestinstitute.eu

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_8

The chapter is based on a review of the documents relating to the preparation of the law on microdata access, as well as on the personal notes and recollections of the author, who headed the research unit of the Finance Ministry and in that capacity was responsible for coordinating the negotiation process that led to the enactment of the law.

2 The Initial Status Quo in Brief

On the eve of its accession to the EU, the conditions for using administrative data for research and policymaking were relatively favourable in Hungary, in terms of the country's bureaucratic traditions and the evolving openness to new public management methods. The accession process had given further impetus to promoting the idea of evidence-based policymaking. At the same time, legal conditions and a low-trust culture in the public sector created considerable barriers.

The Hungarian public sector was built on the foundations of the Austro-Hungarian monarchy[1] and the socialist planned economy,[2] which both involved extensive and systematic data collection and data-based procedures of planning and decision-making in the government (Álló and Molnár 2014). Accordingly, the statistical traditions of the country had also been quite strong. The first official census was held in 1870 and regular household surveys had been carried out since 1949.[3] The Central Statistical Office initiated the introduction of a unique personal identifier in 1978, to facilitate the tracing of individual records over time.

The post-Soviet era brought mixed developments. New, democratic, institutions were established, and citizens' rights against the state were strengthened. While this was a favourable development in general terms, it also involved the creation of barriers to accessing administrative data on citizens. The Constitutional Court abolished the use of the unique personal identifier in 1991. In 1992 a new law on personal data protection was enacted that strongly limited the use of personal information by public authorities and created the position of an Ombudsman to monitor the implementation of the law (Székely 2007).[4] Furthermore, during the

[1] Austrian rule was established after the final defeat of the Turks in 1699 and consolidated in the Austro-Hungarian monarchy between 1867 and 1918.

[2] The bureaucratic procedures of the planned economy followed patterns provided by the Soviet Union, such as the 5-year cycles for industrial targets. The Socialist system was established relatively quickly, starting in 1947, and came to an end in 1989.

[3] The first of these was the household budget survey in 1949, followed by the time budget survey in 1965 and the labour force survey in 1991. The first wage survey that collected individual-level information on wages was conducted in 1986. On the evolution of data protection in the Hungarian census, see Lakatos (2000).

[4] Hungary was among the first post-socialist states to introduce legislation to ensure freedom of information and personal data protection and has become a model for other countries in Central and Eastern Europe (Majtényi 2002; Székely 2007).

1990s, a series of media scandals on the abuse of personal data by public officials increased public mistrust of how the government uses information about citizens and generated fears within the public sector about the potential implications of using individual-level data. During the mid-1990s, there were a few attempts to link administrative databases to identify free-riders of the welfare system. When these failed, mainly as a result of opposition from the Ombudsman and the Constitutional Court, improving access to administrative data began to look like a hopeless endeavour (DPC 1997).

The legal barriers were quite strong. The 1992 law on personal data protection and public information defined data access in such extreme terms that anonymisation for research purposes was legally impossible.[5] On the one hand, public information (more precisely, data of public interest) can be *accessed by all* and anonymised administrative data are considered public information. On the other hand, data are considered personal as long as they can be traced back to the person it refers to, without any flexibility in interpretation.[6] Personal data can be processed only if approved by the person they relate to or if their use fulfils a legal obligation (i.e. use for an official purpose, explicitly stated by law). Until 2007, this implied that data owners had no legal basis for processing personal data for the purpose of anonymisation, since supporting research or statistical analysis was not a legal obligation.

Political developments created somewhat more favourable conditions around 2002–2004. A Socialist–Liberal coalition government was elected in 2002, which was keen to fulfil criteria for Hungary's accession to the EU and had public sector reform on its agenda. In 2003, the Finance Minister established a new research unit within the ministry to strengthen the knowledge base of government decisions. Led by London School of Economics and Political Science graduate Orsolya Lelkes, this unit had some experience in how evidence-based policymaking was implemented in advanced European democracies and also had the necessary skills to apply these methods in Hungary. As shown in the following sections, their efforts finally led to the creation of a new law promoting access to administrative data and new opportunities for creating rich databases by linking several data sources.

[5] See Sect. 3 of Act LXIII of 1992 On the Protection of Personal Data and the Publicity of Data of Public Interest and Majtényi (2002) for a summary of the Act. This Act was replaced by a new law on the same subject in 2011, which slightly eased the requirements for anonymisation, stating that data are regarded personal as long as the connection between the person and the information relating to them is restorable, and this is the case if the data owner is technically equipped for restoring the connection between them (compare Section 4(3) of Act CXII of 2011 on the Right of Informational Self-Determination and on Freedom of Information). Act LXIII of 2012 on the Re-Use of Public Sector Information further strengthened the data access rights of citizens.

[6] Compare this with the UK's rules for research data, which require data owners only to ensure that the probability of re-identification is 'remote'. For more detail on the definition of data of public interest, see Majtényi (2011).

3 Stakeholder Interests

The Finance Ministry had a natural interest in promoting evidence-based policymaking, as its main goal was the efficient allocation of public resources and the curbing of excess spending. The interests of its new research unit, which started to campaign for access to administrative data, largely stemmed from the professional identity of its staff: they had joined the ministry with the aim of promoting evidence-based policymaking, they had the skills to use large-scale administrative data[7] and they understood the potential in accessing the rich data sources of the government.

After a few unsuccessful attempts to acquire administrative data,[8] the Finance Ministry research unit decided to try and remove the legal barriers. To find allies and promote the importance of data access, they held a workshop for data owners and initiated bilateral discussions with potential stakeholders. In 2004, they made a first, poorly prepared attempt to amend the personal data protection law, which failed on the opposition of the Ministry of Justice.

Stakeholders included a wide range of data owners,[9] potential data users (analysts in the civil service and researchers), the Ombudsman for data protection, Neumann Kht (the agency with the IT infrastructure for linking large datasets) and advocacy organisations with an interest in access to public information and personal data protection (Eötvös Intézet and TASZ).

Several stakeholders, such as the Ministry of Education, opposed the new law, fearing that answering data requests would require substantial staff effort and computer time and overburden public institutions. In some cases, data owners may also have feared that external users might discover unlawful or corrupt practices in their institution. Lack of information or trust in anonymisation techniques at the executive level also added to such fears. In more general terms, the lack of trust within and between public institutions (inherited from the socialist regime) also played a role. In such a low-trust environment, any initiative not clearly linked to an interest that all players understand is likely to be viewed with suspicion, on the assumption that the initiator has a hidden agenda.

[7] Prior to the establishment of this unit, there was no expert in the ministry who regularly used econometric models and software in their work. The modelling departments of the ministry did not hold a licence for any statistical software other than Microsoft Excel. On the overall quality of government policymaking, see Verheijen (2007).

[8] The unit first requested and received anonymised extracts of personal tax files from the Tax Authority in 2004. This was relatively easy as the Tax Authority was subordinated to the Finance Ministry, and the request did not involve a linking of several data sources. Next, the unit initiated a request with a plan to link administrative data from the Tax Authority, the Treasury and the Health and Pension Insurance Funds. This attempt failed.

[9] To list the largest: the Central Statistical Office, the Tax Authority, the Health Insurance Fund, the Pension Insurance Fund, the Treasury, the National Labour Office, the Land Registry, the National Railway and thousands of schools and municipalities across the country.

The Central Statistical Office (CSO) strongly opposed the notion that the new law should apply to its data as well.[10] According to its official statement, its main concern was the reliability of the anonymisation process. The CSO feared that the draft law did not provide sufficient guarantees that anonymisation would be complete and that, if this was the case, the CSO would not be able to guarantee anonymity to survey respondents, which in turn might increase non-response rates. Expert-level meetings with the CSO revealed further, possibly more genuine, concerns. First, some CSO officials may have been worried about losing their monopoly on publishing (or selling) the data and losing some revenues. A related issue was the image of the CSO as a reliable source of statistical information. Access to microdata would allow users to publish aggregate statistics of particular variables which may or may not be exactly the same as those published by the CSO. Arguably, if the average journalist or citizen has little knowledge of the intricacies of statistical aggregation and the long list of legitimate causes for such discrepancies, such unofficial statistics might damage the public image of the CSO. Lastly, though the CSO had achieved high professional standards and the general quality of its data was high, some CSO officials may have been concerned that external users would discover some shortcomings in data quality.

Most opponents doubted the need for the new legislation, not being aware of new developments in statistical methods and the potential in using individual-level data.

Though the Gyurcsány government (2006–2008) was broadly supportive of the 'new public management' (NPM) approach, actual demand for evidence-based policymaking remained limited, and thus there was no consensus over the need for improving data access. However, some stakeholders supported the new law because of their commitment to improving policymaking. One of the main supporters was the National Development Agency, which was responsible for allocating EU structural funds and was thus directly exposed to EU expectations to use these funds effectively. Furthermore, as a newly established institution, its staff tended to be better equipped with technical skills and more open to new public management ideas than most of the traditional ministries. Other strong supporters included the State Reform Committee and the Ministry of Economy. An advisor of the Prime Minister's cabinet also actively lobbied for the initiative, as he recognised its potential both for research and for evidence-based policymaking.

Somewhat surprisingly, the Ombudsman for data protection did not raise any serious concerns during the official negotiation process (DPC 2007). There were two likely reasons for this. First, the Ombudsman's mandate covered the protection of citizens' right to information, and he delegated the discussion of the draft law to the unit responsible for this topic. The lawyers in this unit were equally committed to promoting access to data and to the protection of personal data. Second, anticipating their opposition, the Finance Ministry research unit initiated informal negotiations

[10] Letter from Péter Pukli (President of the CSO) to Miklós Tátra (Vice Secretary of State to the Ministry of Finance) on their official comments to the draft law, dated 27 March 2007, Ref. No. 741-77/2/2007-T, 4791/2007.

with the Ombudsman's office before the official process began and, following lawyers' advice, made adjustments to the draft before it was officially submitted for consultation.

4 Negotiation Process Leading to the Law on Accessing Microdata for Policy-Related Analysis

The negotiation process took about a year (see summary in Table 1 below). In the first phase, lasting about 5 months, the Finance Ministry research unit submitted the first draft of the law[11] for comments by the relevant departments within the Finance Ministry and in the meantime initiated informal negotiations with some of the potential opponents. As already mentioned above, discussions with the Ombudsman's office were successful, while the CSO remained sceptical and did not commit to supporting the draft law.

In the second phase, starting in early December 2006, the draft law was submitted by the Finance Ministry for consultation by other ministries and government bodies.

Table 1 Timeline of negotiations on the draft law

Date	Action
May 2006	Concept for the draft law completed
June–November 2006	Formal negotiations within the Ministry, informal negotiations with stakeholders, expert consultations with other ministries
December 2006	Finance Ministry endorses proposal and submits it for cross-ministerial consultations; comments from main data owners and ministries[a]
March 2007	High-level expert meeting[b]
April 2007	Draft law discussed and accepted at the meeting of state secretaries and government
May–June	Draft law discussed in parliamentary committees
June 2007	Law passed by Hungarian parliament
December 2007	Implementation rules enacted

[a] The proposal was sent to 11 ministries, 4 major data owners, the Ombudsman for data protection and 5 public agencies that were potential data users (the Audit Office, the Development Agency, the State Reform Committee, the Innovation Office and the National Academy of Science)

[b] This meeting (*szakmapolitikai értekezlet*) served as a forum for high-level experts in public bodies to discuss draft laws

Sources: Online document archive of the Parliament of Hungary on 'T/3029 *A döntéselőkészítéshez szükséges adatok hozzáférhetőségének biztosításáról*'

[11] The first version of the legislative text was prepared by Máté Szabó, a lawyer, and commissioned by the methodology working group of a ministerial committee for social inclusion, which supported the initiative on the hope that it would lead to better data on income redistribution (Bánfalvi et al. 2006; Szabó 2006). The version in force at the time of writing is available at https://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=a0700101.tv.

After minor amendments, it was passed by the government in April 2007. It was finally enacted by the Hungarian parliament in June 2007 as Act CI of 2007. In this phase, the first setback was posed by the Ministry of Justice, which accepted the purpose and concept of the draft law but disagreed with the proposed legislative text. They took almost 2 months to prepare an alternative solution, after a series of gentle reminders and personal phone calls by high-ranking officials. The second setback came from the Ministry of Education, which fiercely opposed the draft law, fearing that, once implemented, it would impose a large burden on schools. Minor gestures (e.g. further restrictions on who could request data) did not win the Ministry's approval, so in the second meeting of secretaries of state, the Ministry of Education had to be voted down by supportive ministers. As the latter were from traditionally strong ministries (finance and economy), the draft law was safely passed.[12]

Once accepted by ministries, the draft law was easily accepted by the government and did not meet much opposition in the parliamentary committees.[13] It was enacted as Act CI of 2007 by parliament without any amendments or discussion.

5 Reconciling the Requirements of Data Protection and Research

The new law eliminated the main barrier in the preceding legislation by establishing a legal basis for data owners to process personal data for the purposes of anonymisation. This adjusted the balance between meeting the data needs of evidence-based policymaking and those of personal data protection.

During the negotiation process some compromise had to be made to satisfy opponents. In particular, to win the support of *data owners,* the right to request data anonymisation was restricted to public bodies and made to vary by complexity of request. As a result, a request for a highly complex data linkage can be submitted only by a high-level government official, such as a minister or the President of the National Academy of Science (Act CI of 2007).

To conform with the strict standards set by personal data protection rules, the law prescribes a complicated linking procedure based on irreversible identifiers[14] and introduces a number of explicit and (in some aspects, overly) strict rules for anonymisation. These include the restrictions that no sample can be larger than

[12] The fact that the law was proposed by the Finance Ministry was instrumental in the relatively smooth enactment process; the Finance Ministry had a strong position due to its role in allocating resources across ministries. Also it typically proposed fewer new laws than most other ministries, which increased the significance of opposing any of its proposals.

[13] The draft was discussed in four committees without significant opposition, though delegates of the opposition parties did not formally support it in the voting procedure (for details, see archive of the Parliament of Hungary, T/3029).

[14] These are called hash codes, which are unique identifiers but cannot be traced back to the original person.

50% and that geographical identifiers cannot be more detailed than small region level. Furthermore, the data owners must delete both the code and the data soon after sending them to the intermediary agency responsible for linking the data. The intermediary agency is also obliged to keep track of the accumulation of data by owners and maintain a searchable public database of anonymised datasets for secondary use.

6 The Impact of the Law on Microdata-Based Social Science Research

The impact of the new law has not been systematically documented. There is no information available on simple requests when users obtain anonymised data from a single data owner. Requests for linking datasets can be traced as the government agency responsible for linking anonymised datasets (initially Neumann Kht, and since 2011 the National Infocommunications Service Company (NISZ)) is obliged to report on its activities. According to their records, the first linked dataset that was created with reference to the new law was completed in 2010. Since then, around one to three linked datasets have been created every year (see the Appendix for more detail). Some of these involve only two data owners, while the largest involves six or seven public institutions. Ironically, one of the few agencies that has filed several requests is Educatio, an agency established by the Ministry of Education (a former opponent of the law), which uses the linking facility to track the labour market performance of university students after graduation (Nyüsti and Veroszta 2014).

The Institute of Economics of the Hungarian Academy of Science has also made several data requests and has invested substantially in establishing a store of systematically cleaned datasets, which includes several linked databases.[15] These have been widely used by Hungarian researchers and have augmented the publication performance of the institute.[16] The use of administrative data has also contributed to the accumulation of policy evidence, e.g. on the effectiveness of active labour market policies, the income effects of tax cuts or the disadvantages faced by Romani school children (Köllő and Scharle 2016). It should be noted, though, that the use of administrative data has not permeated into the government decision-making process. Clearly, the improved availability of data is a necessary but not sufficient condition for introducing evidence-based policymaking.[17]

[15] For details, see the website of the Hungarian Academy of Sciences, Institute of Economics (IE) Databank at http://adatbank.krtk.mta.hu/nyito.

[16] Since 2007, several papers using linked administrative data from Hungary have been published in high-ranking journals such as the *American Economic Review* (e.g. Kertesi and Kézdi 2011; Halpern et al. 2015).

[17] Though an agency (ECOSTAT Kormányzati Hatásvizsgálati Központ) was established in February 2011 to prepare (and support ministries in preparing or subcontracting) impact evaluations, it

7 Summary and Conclusions

Hungary introduced a law ensuring access to anonymised personal data for research and policymaking in 2007. The law has forged a compromise between strict provisions on personal data protection and researchers' needs for microdata that has passed the test of practical application. The Hungarian case may be a model for improving access to administrative data for research and policy analysis in a low-trust environment.

The present review of the process leading to the enactment of the law highlighted three notable enabling factors. First, it was important to have a credible and dedicated insider, in a strong ministry, who could invest the time and effort in lobbying and coordination. This went together with the general though vague support of the government for improving the evidence base of policymaking. Second, early negotiations with influential stakeholders such as the Data Protection Commissioner and the involvement of potential supporters such as the National Development Agency seem crucial for smoothing the formal negotiation process.

Lastly, though the law has several weak points, it has enabled not only the creation of rich datasets but also the accumulation of experience, thus reducing ignorance-based attitudinal barriers (regarding the need for individual data and anonymisation methods) and fears about possible misuse by researchers. This will facilitate negotiating the necessary corrections to the law when demand for evidence-based policymaking revives.

Acknowledgements Invaluable comments and suggestions by Csaba Gáli (who prepared the final draft of Act CI of 2007 as a civil servant in the Ministry of Justice) and Balázs Váradi (who was an advisor to the Prime Minister in 2007 and supported its enactment in that capacity) are gratefully acknowledged.

Appendix: Completed data-linking procedures between 2007 and 2015

was merged into another agency and lost most of its powers in July 2012 (Government Decree 177/2012).

Year	Procedures (data owners involved[a])	Institutions requesting linked data	Data owners involved								
			National Tax Authority	National Labour Office (wages)	National Labour Office (Unemployment)	Treasury	Pension Insurance Fund	Health Insurance Fund	Education Education Office[b]	Central Statistical Office	Student Loan Centre
2008	0	n.a.									
2009	1 (2)	Finance Ministry	x	x							
2010	2 (2, 4)	Institute of Economics HAS	x	x	x	x	x	x			
2011	3 (2, 3, 4)	Development Agency, Educatio[a]	xx		xxx	x		xx	x		
2012	1 (6)	Ministry of Economy	x		x	x	x	x		x	
2013	1 (3)	Development Agency	x		x	x					
2014	3 (2, 6, 7)	Educatio	xx		xx	xx	xx	xx	xxx		xx
2015	2 (2, 5)	Institute of Economics HAS	xx	x	x		x	x	x		
2016	2 (6, 6)	Educational Authority	xx		xx		xx	xx	xx		xx

[a] The numbers in brackets indicate the number of data owners for each data-linking procedure initiated during the given year. The crosses in the table show which data owners were involved

[b] Educatio Kht is an agency established by the Education Ministry for implementing education reforms

HAS Hungarian Academy of Science, *n.a.* not applicable

Source: http://www.nisz.hu/hu/kozadat accessed on 28 January 2017 and further details provided on request by the National Infocommunications Service Company (NISZ)

References

Act CI of 2007 (2007) évi CI. törvény a döntéselőkészítéshez szükséges adatok hozzáférhetőségének biztosításáról (Act CI of 2007 on the provision of access to information required for drawing up decisions). https://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=a0700101.tv. Accessed 5 Jun 2017

Álló G, Molnár S (eds) (2014) A 'hiteles helyektől' az elektronikus közigazgatásig. Mérföldkövek a hazai közigazgatás automatizálásának és a kormányzati számítástechnika kialakulásának történetében. Primaware, Szeged http://mek.oszk.hu/14600/14683/14683.pdf. Accessed 31 Aug 2017

Bánfalvi I, Havasi É, Ulicska L (2006) Bemutatkozik a METOD az ICSSZEM égisze alatt működő Társadalmi Kirekesztődés elleni Bizottság Módszertani Műhelyének céljáról és tevékenységéről – A METOD megalakításának szempontjai, a Műhely célja, legfontosabb alapelvei. Esély 2006(3):25–29 http://www.esely.org/kiadvanyok/2006_3/BANFALVI.pdf. Accessed 5 Jun 2017

DPC (1997) Beszámolók IV. Vizsgálatok (Annual report of the Data Protection Commissioner). http://abi.atlatszo.hu/index.php?menu=beszamolok/1997/IV/A/1/10. Accessed 5 Jun 2017

DPC (2007) Adatvédelmi biztos állásfoglalása a 2007/CI törvény tervezetének parlamenti vitájához. Ügyszám: 472/J/2007-11. http://abi.atlatszo.hu/index.php?menu=adatvedelem&dok=472_J_2007-11. Accessed 5 Jun 2017

Halpern L, Koren M, Szeidl Á (2015) Imported inputs and productivity. Am Econ Rev 105(12):3660–3703

Kertesi G, Kézdi G (2011) The Roma/non-Roma test score gap in Hungary. Am Econ Rev 101(3):519–525

Köllő J, Scharle Á (2016) Data revolution in Hungary: clerks, cronies and the chosen few. Presentation at a CRIE workshop, Centre for Research and Impact Evaluation, Ispra, Italy, 8 March 2016. https://crie.jrc.ec.europa.eu/?q=content/%C3%A1gota-scharle-budapest-institute-policy-analysis-and-j%C3%A1nos-k%C3%B6ll%C5%91-institute-economics. Accessed 5 Jun 2017

Lakatos M (2000) Az adatvédelem jogi szabályozása a magyar népszámlálások történetében. Statisztikai szemle. http://www.ksh.hu/statszemle_archive/2000/2000_10-11/2000_10-11_794.pdf. Accessed 5 Jun 2017

Majtényi L (2002) Ensuring data protection in East-Central Europe. Soc Res 69(1):151–176 www.jstor.org/stable/40971541. Accessed 5 Jun 2017

Majtényi L (2011) Freedom of information, disclosure, privacy, and secrets in Hungarian law. Eur Integr Stud 9(1):73–85 http://www.matarka.hu/koz/ISSN_1588-6735/GTK_vol_9_no_1_2011_eng/ISSN_1588-6735_vol_9_no1_2011_eng_073-085.pdf. Accessed 5 Jun 2017

Nyüsti S, Veroszta Z (2014) Hungarian graduate career tracking 2013: integration of administrative databases. Educatio, Budapest https://www.felvi.hu/pub_bin/dload/DPR_tanulmanyok/dpr_integration_of_data_2013_en_VEGLEGES_web.pdf. Accessed 5 Jun 2017

Szabó MD (2006) Hozzáférés adminisztratív és statisztikai egyedi adatokhoz empirikus vizsgálatok végzése céljából. Készült a METOD megbízásából, 2006. Manuscript, May 2006. Budapest

Székely I (2007) Central and Eastern Europe: starting from scratch. In: Florini A (ed) The right to know: transparency for an open world. Columbia University Press, New York, pp 116–142

Verheijen T (2007) Administrative capacity in the new Member States: the limits of innovation? World Bank Working Paper 115. World Bank, Washington, DC

Ágota Scharle is a labor economist who earned a Master of Economics at Corvinus University, Budapest, in 1994, and a PhD in Economics at the University of Oxford in 2001. She is now a senior researcher at the Budapest Institute for Policy Analysis, an independent think tank based in Budapest, Hungary. Her recent work has focused on the impact evaluation and design of social

policy and active labor market policies for disadvantaged jobseekers, and the political economy of welfare reforms. She was Head of research in the Hungarian Finance Ministry between 2005 and 2008, had worked as an economist at the Finance Ministry (2003–2005) and at the National Employment Office (1994–1996).

Experimental and Longitudinal Data for Scientific and Policy Research: Open Access to Data Collected in the Longitudinal Internet Studies for the Social Sciences (LISS) Panel

Marcel Das and Marike Knoef

1 Introduction

Empirical research needs data. Scientists are able to collect data through small-scale experiments in laboratories and in many cases by using students as respondents. There are also successful initiatives to collect new data in large-scale longitudinal surveys. The longest-running longitudinal household survey is the Panel Study of Income Dynamics (PSID; https://psidonline.isr.umich.edu/). Other examples include the Health and Retirement Study (HRS); the Survey of Health, Ageing and Retirement in Europe (SHARE); Understanding Society (incorporating the British Household Panel Survey); and the German Socio-Economic Panel Study (SOEP). Access to data is in most cases entirely open, subject to signing a statement governing the use of the data.

Empirical researchers benefit from an open-access policy. They have access to a huge number of datasets. However, there are also some problems. Data collection, as in the above examples, is carried out mostly by face-to-face interviewing. This is rather time-consuming, and public release of the data often takes place more than a year after data collection has been completed. In addition, there is little or no room for "outsiders" to add new questions or experimental modules. And although the longitudinal surveys have a multidisciplinary setup, possible new questions must fit the context of the survey. Finally, the large-scale longitudinal surveys all suffer from the problem that face-to-face interviewing has become extremely expensive.

M. Das (✉)
CentERdata and Tilburg University, Tilburg, Netherlands
e-mail: das@uvt.nl

M. Knoef
Leiden University and Netspar, Tilburg, Netherlands
e-mail: m.g.knoef@law.leidenuniv.nl

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_9

As an alternative, a new initiative was started in the Netherlands a decade ago: the Longitudinal Internet Studies for the Social sciences (LISS) panel. With a major investment by the Dutch government, a large-scale infrastructure was developed for the use of social science scholars (and others). The panel is administered by CentERdata, a nonprofit research organization housed at the campus of Tilburg University. The infrastructure is entirely open access; only commercial use of the data is prohibited. Researchers from all disciplines are invited to submit their (longitudinal) surveys and experiments. Due to the setup of the infrastructure, the data can be released quickly in comparison with the large-scale longitudinal surveys mentioned above.

The LISS panel was built in close collaboration with Statistics Netherlands (SN), the Dutch statistical office. SN drew a random sample of addresses from the population register, and all selected households were contacted in a traditional way, either by telephone or in person. At the end of the recruitment interview, the household respondents were asked whether or not they were willing to participate in an online panel. If there was no computer and/or Internet connection in the household, CentERdata provided the necessary equipment. All household members aged 16 years and older were asked to participate. If at least one household member agreed to participate, the household was included in the panel.

Researchers are charged a fee to use the infrastructure to collect new data. However, once the data have been collected, access to the data archive is free of charge. More information about the LISS panel, including the setup, can be found in Scherpenzeel and Das (2011).

The aim of this chapter is to explain the LISS infrastructure and the opportunities it offers for data-based policy research. The remainder of the chapter is organized as follows. Section 2 describes the types of data collected in the LISS panel, the innovations in data collection which were tested in LISS, the open-access data policy, and the option to link survey data to administrative resources available at SN. Section 3 gives some examples of how the LISS infrastructure has been used to study policy-relevant issues. Finally, Sect. 4 concludes with some remarks on future developments and challenges.

2 The LISS Infrastructure and Linkage with Other Sources

2.1 Longitudinal Core Study

To prevent panel members having to answer similar questions month after month—due to the popularity of certain topics at certain times—it was decided to have a rich and lengthy core questionnaire. This core questionnaire, designed with assistance from international experts in the relevant fields, follows changes over the life course of individuals and households. The questionnaire is repeated annually and covers eight modules, each with its own theme:

- Health
- Politics and values
- Religion and ethnicity
- Social integration and leisure
- Family and household
- Work and schooling
- Personality
- Economic situation: assets, income, and housing

Data from the longitudinal core study allow for analyses of changes in people's lives, their reaction to life events, and the effects of societal changes and policy measures. For example, the core modules "'Politics and Values" and "'Social Integration and Leisure" were used by Van Ingen and Van der Meer (2016) to test four possible explanations for the well-documented correlation between civic engagement and political socialization. Kalmijn (2015) used data from the core module on "Family and Household" to examine the effects of divorce and re-partnering on the relationships that fathers have with their adult children. Six waves of the "Health" module were used by Cabus et al. (2016) to estimate the short-run causal effect of tumor detection and treatment on psychosocial well-being, work, and income.

The major strength of the longitudinal core study, however, is the opportunity it provides to combine data from studies and experiments proposed by researchers with data from the core study. This greatly enhances the cost-effectiveness and scientific value of the experimental modules proposed by researchers. It eliminates the need to collect an array of background variables in each survey or experiment, and allows for links with a wealth of other (non-retrospective) information available on the panel members.

An example is a multi-wave study on mental health (Lamers et al. 2011). The researchers who proposed this study argued that there is a growing consensus that mental health is not merely the absence of mental illness but also includes the presence of positive feelings (emotional well-being) and positive functioning in both individual life (psychological well-being) and community life (social well-being). Lamers et al. examined a new self-report questionnaire for positive mental health assessment (the so-called Mental Health Continuum-Short Form (MHC-SF)). The collected data were enriched with data from the longitudinal core study (modules "Health," "Personality," "Politics and Values," and "Social Integration and Leisure"). So far, data from this particular multi-wave study combined with data from the data archive have resulted in ten articles published in peer-reviewed scientific journals, a book chapter, a Master's thesis, and a PhD thesis. In addition to the scientific value, the project has also had a societal impact, as the MHC-SF is now widely used by the Dutch Association of Mental Health and Addiction Care.

Another example in which data were successfully merged with the longitudinal core study is a project that aims to determine whether or not people change their reform preferences when faced with increasing reform pressures such as an aging society (Naumann et al. 2015). The researchers collected data in July 2013, September 2013, and January 2014 and combined their data with data from the core modules "Economic Situation (Income)," "Politics and Values," and "Work and Schooling" from the 2008 and 2013 waves. Naumann et al. confirmed theoretical expectations that people change their support for unemployment benefits in reaction to changes in their individual material circumstances. Job loss leads to increased support for public unemployment benefits. The availability of longitudinal data, in particular, covering the period of the international economic crisis (2008–2009), made the analysis possible.

2.2 Experimental Data

The LISS panel has also been used successfully for various types of experiments. Before the main recruitment of the LISS panel even started, a comprehensive pilot study was fielded to determine the optimal recruitment strategy for an infrastructure such as LISS (Scherpenzeel and Toepoel 2012). The factors that were considered in the pilot study were contact mode (recruitment either by telephone, in person, or by a combination of these methods), incentive amount, timing of the incentive, content of the advance letter, and timing of the panel participation request. Scherpenzeel and Toepoel showed that all incentives were found to have much stronger effects on response rates when they were distributed with the advance letter (prepaid) than when they were paid later (promised). The highest response rate was found with a prepaid incentive of EUR 10. For more results of the recruitment pilot, we refer to Scherpenzeel and Toepoel (2012).

The LISS panel is an ideal infrastructure for studying survey methodological issues, in particular, when they relate to the mode of interviewing (online). However, many experiments contributing to substantive research were also run in LISS. For example, Bellemare and Sebald (2011) presented a class of two-player extensive-form games allowing measurement of belief-dependent preferences, including guilt aversion as an important special case. A total of 2000 LISS panel members were invited to participate in a sequential game that was played across 2 consecutive months. Bellemare and Sebald found evidence of significant guilt aversion in the Dutch population: a significant proportion of the population was found to be willing to pay to avoid letting down the other player, in line with predictions of belief-dependent models of guilt aversion.

Another example of a substantive experiment concerned ethical behavior and class status (Trautmann et al. 2013). In this experiment, randomly selected LISS panel members had to make decisions that determined how much money they and someone else would earn. Trautmann et al. showed that ethical behavior is affected by moral values, social orientation, and the costs and benefits of taking various actions. Strong class differences emerged in each of these areas, leading to differences in behavior.

2.3 Innovations in Data Collection

One of the goals in LISS is to innovate data collection methods through experiments with new technologies. The first large-scale experiment started in 2010. A random sample of (about) 1000 LISS households was provided with an advanced bathroom scale. This scale measures body weight and impedance (on which fat and muscle percentage is based). The scale establishes a wireless connection with the gateway via a radio signal. This minimizes the respondents' burden to provide the data; they are requested only to step on the scale (once a day, once a week, or at an unspecified frequency). A first empirical analysis is reported in Kooreman and Scherpenzeel (2014), based on almost 80,000 measurements collected in 2011.

The measurement of time use typically relies on paper diaries. As this is quite burdensome for the respondents, response rates in time-use surveys are generally low. In addition, the traditional setup is rather expensive. A smartphone allows time-use data to be collected in a more efficient way. In close collaboration with the Netherlands Institute for Social Research (SCP), a pilot study was carried out in the LISS panel to test the feasibility of collecting time-use data with smartphones. A total of 2000 LISS panel members participated in the study; some members of the sample did not own a smartphone and were lent one for a short period of time. A time-use app was developed specifically for this study, which had similarities with the paper version. There were also differences with the paper version, such as the possibility of copying repeated activities from a previous time slot and of filling in activities such as sleeping and working for longer time periods. The results of this feasibility study can be found in Sonck and Fernee (2013).

Smartphones can also be used to track travel behavior, using the phone's global positioning system (GPS) functionality. Traditionally, data on travel behavior are collected through cross-sectional travel surveys using a paper diary. This entails a serious time investment by the respondent, especially when travel data are collected for a longer period of time. Moreover, due to memory effects, the accuracy of the data is rather low. Using a dedicated app, travel data were collected from a random selection of LISS panel members. Data collection took place in 3 years (2013, 2014, and 2015); in each year a random selection of 500 panel members participated in the study, using their own smartphone with the app installed or a loan smartphone. Geurs et al. (2015) analyzed the first dataset and concluded that using the app is a promising alternative to traditional travel diaries.

The measurement of physical activity is a final example of an experiment with a new measurement device. The goal of this experiment was to form a more realistic and complete picture of physical activity when objective measures and self-reports are combined, particularly in the context of international studies on physical activity. The study involved an accelerometer, developed by GENEActiv (https://www.activinsights.com/products/geneactiv/). The device is wearable as a watch and is waterproof. It measures acceleration in three dimensions, body temperature, and light intensity, at a frequency of 60 measurements per second (60 Hz). Approximately 1000 LISS panel members participated in the main study, using 300 devices. Panel members wore the device for 8 consecutive days, day and night. For each participant this resulted in a large dataset; for the entire sample in the experiment, approximately 5 terabytes of raw data were produced. The same devices were shared with research teams running the English Longitudinal Study of Ageing (United Kingdom) and the Understanding America Study (United States). Interesting results were obtained. Kapteyn et al. (2018) showed that self-reports and objective measures of physical activity tell a strikingly different story about differences between the Netherlands and the United States: for the same level of self-reported activity, the Dutch are significantly more physically active than Americans.

2.4 Open-Access Data Policy

Access to the data collected in LISS is open to every researcher, free of charge, both in the Netherlands and abroad. Data are made available through the advanced LISS Data Archive (www.dataarchive.lissdata.nl/). This archive, based on existing international specifications, has been awarded the Data Seal of Approval (www.datasealofapproval.org), confirming adherence to the guidelines for trusted digital repositories. Any researcher who signs a confidentiality statement can use the data. Use of variables collected in different waves (or studies) is facilitated by allowing researchers to collect such variables in a "shopping basket," which then automatically generates a dataset according to the user's specification(s).

In June 2018 more than 2800 users were registered, affiliated with more than 100 institutes worldwide (including top universities such as Harvard, Stanford, and the University of Michigan). Data are used for both scientific and policy-relevant/socially relevant research. So far, more than 491 papers based on LISS data have been published, including 236 articles in peer-reviewed international scientific journals and 31 PhD theses.

2.5 Linking to Administrative Data

SN collects a large variety of data which can be accessed by researchers under very strict privacy and confidentiality procedures. Through collaboration between SN and CentERdata, all LISS data can be linked to administrative data. This is only possible within the remote access environment of SN. Researchers can send LISS data to SN through a secure connection. SN then matches identification numbers from LISS panel members with the identification numbers available in SN's records. LISS panel members are informed about this linking and can opt out at any time. Once a particular person has opted out, record linkage for that person is no longer possible. Less than 10% of panel members have opted out. Linkage with administrative data lays the groundwork for an even richer data resource, as it is possible to augment survey records with administrative data on, for example, labor, income, wealth, pension entitlement, and health care. Section 3 presents some examples of research projects based on LISS data combined with data from SN registers.

Some studies (pending publication) link LISS data to external data sources available from institutes other than SN. One study links data from the core study ("Health" module) to data on air pollution (available from the Dutch National Institute for Public Health and the Environment (RIVM)). A second example is a study that links LISS data to weather data (available from the Royal Netherlands Meteorological Institute (KNMI)). In both cases files with postal codes and the variable of interest are merged to create a file with LISS data enriched with postal codes. The merging of files is performed by CentERdata; the researcher receives the merged file excluding the postal codes. As long as the variable of interest is at a sufficiently high level of aggregation, no individual panel member can be identified in this way.

3 Use of LISS Data for Policy-Relevant Research

3.1 Societal Challenges

Society is currently experiencing a number of significant trends, with a host of attendant challenges:

- The population is aging, with implications for the cost of health care, sustainability of the social security and pension systems, and the structure of labor and product markets.
- Health disparities are substantial, and health status varies strikingly by socioeconomic status. Obesity rates are rising sharply, and levels of physical activity are falling.
- The immigrant population has become sizeable and has not been effectively integrated into the labor market, the educational system, or the social fabric.
- Volatility in the international financial system has put savers' investments at risk.
- Work patterns are changing across generations and age groups, as well as by gender. With people now working until later in life, older workers often require adaptations in workplaces and shorter working hours. Female participation in the labor force has increased dramatically, but women are much more likely to work part time than men, with implications for career prospects and compensation.

Through its open-access data policy, LISS offers a rich and valuable source of information that can be used to address the challenges posed by these and other trends. Sound policy that can positively shape the future of citizens will depend on high-quality research in the social sciences to inform decision-making, both in government and in industry.

Survey data from LISS in combination with administrative data have played an important role in policy discussions on the future of the Dutch pension system. Results have also been used for a “pension coach” app to help the Dutch public prepare for retirement. The next three subsections describe policy-relevant research based on combinations of LISS panel data and administrative data and its impact on society.

3.2 Retirement Savings Adequacy

Population aging and the poor performance of financial markets in recent years have put the sustainability of pension arrangements in many Western countries under pressure. To investigate if the Dutch population will be able to cope with possible cutbacks in pension benefits, De Bresser and Knoef (2015) analyzed their preparedness in 2008, on the eve of the prolonged economic slump. To do so they compared self-reports of minimal and preferred expenditures during retirement with annuitized wealth from administrative data. The rationale for this approach is that preferences and constraints are likely to vary across individuals and households. Measuring readiness against a single universal threshold, such as a retirement income equal to or greater than 70% of previous earnings,[1] fails to capture relevant differences in coping strategies.

For the subjective assessment of minimal and preferred expenditure levels during retirement, De Bresser and Knoef (2015) used data elicited from the LISS panel in January 2008 on the initiative of Johannes Binswanger and Daniel Schunk (Binswanger et al. 2013). Due to the open-access policy of LISS, these data were still available 5 years later and could be merged with other data from the LISS data archive.

A question about minimal retirement expenditure was raised at the beginning of the survey, after a couple of items regarding housing costs during retirement. The question was phrased as follows:

This question refers to the overall level of spending that applies to you [and your partner/spouse] during retirement. What is the minimal level of monthly spending that you want during retirement?

Please think of all your expenditures, such as food, clothing, housing, insurance, etc. Remember, please assume that prices of the things you spend your money on remain the same in the future as today (i.e., no inflation).

The quality of any evaluation of retirement readiness depends on ability to measure financial resources. Survey reports of assets are known to suffer from substantial non-response and under-reporting, particularly when it comes to categories of ownership such as stocks and savings accounts (Bound et al. 2001; Johansson and Klevmarken 2007). Therefore, De Bresser and Knoef (2015) preferred to use more

[1] This is a widely accepted standard in the literature (Haveman et al. 2007).

reliable administrative sources. They matched the LISS survey data with tax records and data from pension funds and banks that are available at SN. This allowed them to construct a complete and precise measure of the resources available to households.

The quality of the self-reported expected retirement expenditures is also important. This depends on the degree to which people can predict their expenditure needs during retirement. De Bresser and Knoef (2015) showed that people report reasonable expenditures compared with their current income level. Furthermore, young people provide similar answers to retirees, who know what it is like to be retired. Finally, the model controls for the fact that some individuals have thought about retirement more than others and that some people will find it more difficult than others to answer questions about consumption needs during retirement.

De Bresser and Knoef (2015) found that, overall, the Dutch population was well prepared for retirement. The median difference between the after-tax annuity that can be obtained at age 65 and the individual-specific level of minimal expenditure was 25%, taking into consideration public and occupational pensions. Still, for a sizable minority of the sample, close to 20%, the annuity falls short of minimum expenditure, even if all sources of wealth are taken into account (including private savings and housing wealth, in addition to public and private pensions). The size of those deficits is large enough to be problematic, with a median shortfall of around 30%. The self-employed and the divorced stand out as vulnerable groups with relatively modest pension entitlements.

The results of De Bresser and Knoef (2015) stimulated the policy debate in the Netherlands on shortages but also on households that save more than they need to finance their retirement. As explained by Knoef et al. (2015),[2] there is considerable variation in retirement savings adequacy. The results of the study were commented on print media, on the radio, and on televised news programs in the Netherlands. The advisory report by the Social and Economic Council on the future of the Dutch pension system referred to the results (SER 2015). This advisory report was taken seriously by the relevant ministries in their plans to reform the pension system. Furthermore, the results were used by pension funds and insurance companies in the pension field to gain insight into the composition of wealth in Dutch households. The results were also presented to the State Secretary of the Ministry of Social Affairs and Employment, who was especially interested in vulnerable groups (regarding pension accumulation). Finally, the Dutch central bank made use of the results in its paper on Dutch household balance sheets (DNB 2015). With this paper the Dutch central bank aimed to limit disruptive tax incentives with regard to the accumulation of wealth by Dutch households.

[2] This is a policy brief in which the results of De Bresser and Knoef (2015) play an important role, together with those of Knoef et al. (2016).

3.3 Retirement Expenditure Goals After the Crisis

The Dutch Authority for the Financial Markets raised the question of whether or not retirement expenditure goals changed after January 2008, in response to the financial crisis. Therefore, a new questionnaire was fielded in the LISS panel in December 2014, again asking about retirement expenditure goals. In addition to the question on minimal retirement expenditure described above, De Bresser et al. (2018) also asked a question about preferred retirement expenditure.

Figure 1 shows the median of minimal and preferred retirement expenditure goals by income quintile.[3] Both minimal and preferred retirement expenditure goals increase with income. However, when retirement expenditure goals are divided by current income, these ratios decline with income. Whereas the median poor person needs about 100% of current income after retirement, the median rich person needs about 60%.

Descriptive statistics about minimal and preferred expenditure are in themselves interesting. Therefore, they are used by a large Dutch insurance company in its "pension coach" app, which is available for the whole Dutch population free of charge. After filling in retirement expenditure goals and wealth and pension entitlements, the app indicates what the person must do to reach his/her retirement expenditure goal. To help people determine their retirement expenditure goal, they are offered information on the interquartile range of retirement goals that peers with the same income and household situation reported in the representative LISS panel.

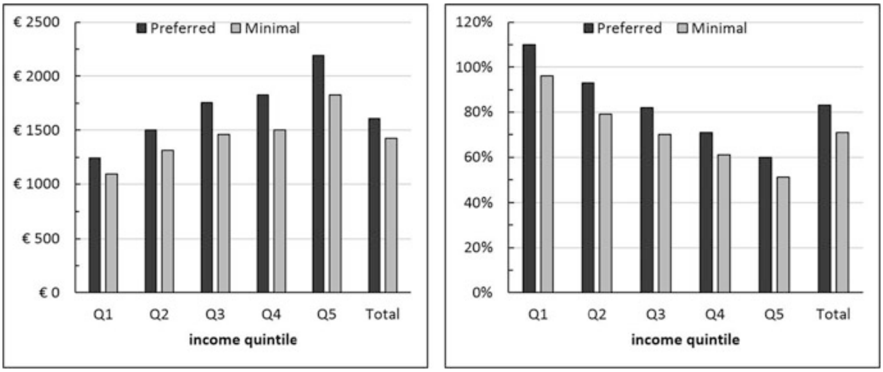


Fig. 1 Retirement expenditure goals by income quintile. Note: This figure shows preferred and minimal retirement expenditure goals (left) and preferred and minimal retirement expenditure goals divided by current income (right). Source: Own calculations based on data described in De Bresser et al. (2018)

[3] Retirement expenditure goals are standardized and divided by standardized net income. Income quintiles are based on standardized gross income.

Comparing the data for January 2008 and December 2014, De Bresser et al. (2018) show that minimal retirement expenditure goals (in real terms) declined by about EUR 200 per month over that period. In particular, high-income individuals, homeowners, widows, and men with self-employed individuals in their households reduced their retirement expenditure goals.

People may have adjusted their pension ambitions downward because of gloomy media reports about pensions. Or, in line with the life cycle model of Modigliani and Brumberg (1954), individuals may smooth out exogenous wealth shocks from the crisis over their remaining life cycle by changing their expenditure and/or labor supply. De Bresser et al. (2018) show that a shock in pension wealth of EUR 100 reduced retirement expenditure goals on average with EUR 23-33. In addition, gloomy media reports may have led to lower expectations regarding the future (for all people, not related to individual declines in expenditure goals).

The results were presented at a World Economic Forum expert meeting and were used by insurance companies to inform their financial advisors on "what is an adequate pension." The Dutch Authority for the Financial Markets and the Ministry of Finance used the results in Fig. 2 to identify groups that are vulnerable in terms of retirement savings adequacy. For the year 2017, the Ministry of Finance has announced a focus on divorced people, based partly on these results.

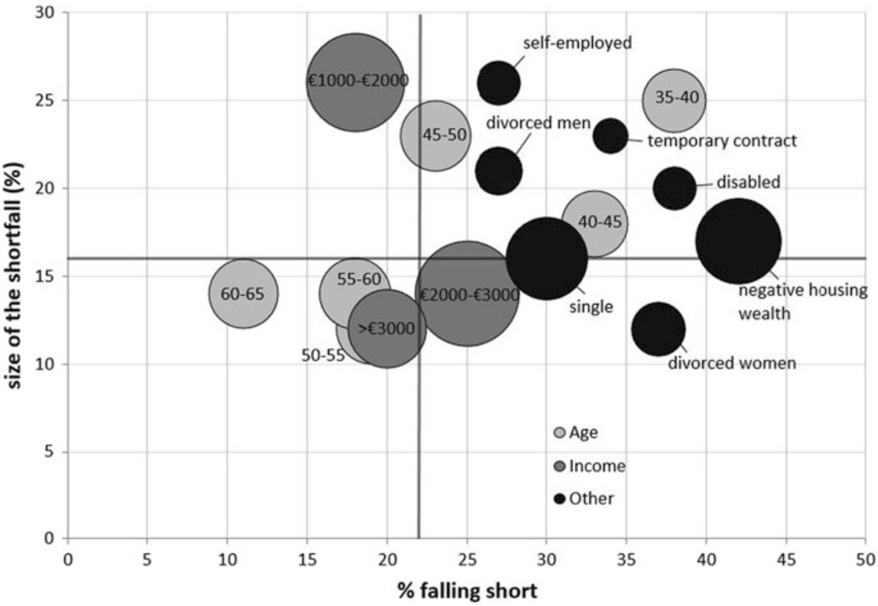


Fig. 2 Retirement savings adequacy in subgroups of the population. Source: Authority for the Financial Markets (2015)

Expected pension annuities are also calculated in Knoef et al. (2017) for a large administrative dataset (the Income Panel Survey, covering about 90 000 Dutch households). Here, as a benchmark to judge savings adequacy, a 70% replacement rate was used (e.g., Haveman et al. 2007) but also income-dependent replacement rates based on the self-reported expenditure goals in the LISS panel (Fig. 1). These results were used by the Ministry of Social Affairs and Employment (SZW 2016) to formulate policy options to reform the Dutch pension system (ahead of the elections in March 2017). Policies were proposed to stimulate pension accumulation for the self-employed (since the results showed the self-employed to be a vulnerable group with regard to pension accumulation). The report also proposed to differentiate between pensions for homeowners and renters, since the results showed large differences in the pension accumulation of renters and homeowners. The Ministry of Finance used the results for its study group on sustainable growth and also in the annual government report that clarifies the expected income and expenditure of the national government.[4]

3.4 Stated Preference Analyses to Guide Policies

Sometimes preferences for policies cannot be measured, for example, because the policy is not yet in place. In such cases, surveys can be used to estimate the behavioral responses of individuals to certain policies by a stated preference analysis. In a stated preference analysis, respondents are typically placed in a hypothetical decision context and are asked to make several choices. In this way preferences can be elicited which can guide policies in the near future. Below the authors describe three stated preference analyses in the LISS panel that guide policies in an aging society. The first two are about the labor market for older workers, and the third is about the design of long-term care.

Kantarci and Van Soest (2013) estimated preferences for retirement plans. By using a stated preferences analysis, they studied the effects of pension incentives and increasing retirement age on the preferences for retiring full time or part time at a later age. They find that two in five respondents prefer partial retirement over early or delayed abrupt full retirement. This suggests scope for policy interventions that emphasize partial retirement plans, offering flexible solutions for employees to optimize their retirement paths. Furthermore, the results show that individuals are responsive to an increasing retirement age at both the extensive and intensive margins.

Oude Mulders et al. (2014) fielded a stated preference analysis among all managers in the LISS panel, to examine a manager's considerations in the decision to rehire employees after mandatory retirement. This information is important for governments that want to increase the labor force participation among the elderly.

[4] Known as the *Miljoenennota*.

The results show that employers are strongly affected by employees who offer to work for a significantly lower wage. Overall, employers are disinclined to rehire employees after mandatory retirement, although large differences exist between employees.

More recently, Van Ooijen et al. (2017) investigated the public's willingness to pay for long-term care in the Netherlands. Their first results show that people are more willing to pay for household chores and personal care than for chaperoning, entertainment, or insurance for the purchase of medical devices. First results also show that people who expect to need relatively more long-term care are more inclined to buy an insurance plan for long-term care. This means that adverse selection is likely to be a serious concern when long-term care responsibilities are transferred from the government to the individual and insurers wish to enter the market.

4 Future Developments and Challenges

This chapter explained the LISS panel, an ultramodern and (cost-)efficient research infrastructure that is now solidly in place. More than 9 years' worth of rich and innovative data has now been collected through this infrastructure. Researchers worldwide have accessed the data for use in scientific, policy, and societal studies. The open data policy has facilitated a vast amount of monodisciplinary research, but multidisciplinary research is also possible by merging data from different disciplines. The authors have shown how results of the LISS panel in combination with administrative data have an influence in the political arena.

The world of primary data collection will change. New forms of data collection, including wearable computing and data collected through sensor technology, will replace the more traditional ways of collecting information. Lengthy surveys might be replaced by shorter high-frequency surveys. Infrastructures such as the LISS panel can provide a natural environment to launch new forms of data collection and to conduct high-frequency data collection. Disseminating the data may become a challenge, however, in part due to the size of databases—as in the accelerometer example described in Sect. 2.3—but also on account of privacy regulations. The more detailed information that becomes available on individuals, the easier it might become to identify individual respondents. Take, for example, the high-frequency data that were collected on travel behavior (Sect. 2.3). The daily GPS data can easily reveal where the respondent lives and works. It is impossible to make these data openly available for the research community, especially in combination with all the other data collected in LISS. All data need to be thoroughly anonymized before they can be made available (e.g., by disseminating only distances travelled).

The biggest challenge concerns the funding of infrastructures such as LISS. It is a public benefit and needs budgets from scientific funding agencies to maintain its high-grade scientific standards. Generating income through users of such infrastructures might be an option, but this may not be to the advantage of the open data policy. Users who pay for their data are, in general, not in favor of immediately sharing their data with others. And even if they are willing to share the data, resources are required to make the data dissemination—including accessible metadata—feasible. It seems unfair to charge these costs to the individual researcher or research team who commissioned the survey. Budget is also needed to make the administrative data more easily accessible for both scientific and policy-relevant research. Investments need to be made in computational power and software tools, and to serve the international research community, all documentation of metadata needs to be made available in English.

We live in a society where policies are improved by insights derived from data. These can be survey data but also administrative data, unstructured data, or a combination of these. As a research community, we need to make sure that data remain as accessible as possible.

References

Authority for the Financial Markets (2015) Neem drempels weg opdat Nederlanders in actie komen voor hun pensioen (How to increase pension awareness?) Netherlands Authority for the Financial Markets, Amsterdam

Bellemare C, Sebald A (2011) Learning about a class of belief-dependent preferences without information on beliefs. IZA Discussion Paper No. 5957. Institute for the Study of Labor, Bonn

Binswanger J, Schunk D, Toepoel V (2013) Panel conditioning in difficult attitudinal questions. Public Opin Q 77:783–797

Bound J, Brown C, Mathiowetz N (2001) Measurement error in survey data. In: Heckman J, Leamer E (eds) Handbook of econometrics, vol 5. North-Holland, Amsterdam, pp 3705–3843

Cabus S, Groot W, Maassen van de Brink H (2016) The short-run causal effect of tumor detection and treatment on psychosocial well-being, work, and income. Eur J Health Econ 17(4):419–433

De Bresser J, Knoef M (2015) Can the Dutch meet their own retirement expenditure goals? Labour Econ 34:100–117

De Bresser J, Knoef M, Kools L (2018) Cutting one's coat according to one's cloth: how did the great recession affect retirement resources and expenditure goals? Netspar discussion paper 05/2018-029

DNB (2015) De Vermogensopbouw van huishoudens: is het beleid in balans? (The wealth accumulation of Dutch households: are policies balanced?). Occasional Studies 13.1. De Nederlandsche Bank, Amsterdam

Geurs K, Thomas T, Bijlsma M et al (2015) Automatic trip and mode detection with MoveSmarter: first results from the Dutch mobile mobility panel. Transp Res Proc 11:247–262

Haveman R, Holden K, Romanov A et al (2007) Assessing the maintenance of savings sufficiency over the first decade of retirement. Int Tax Public Financ 14(4):481–502

Johansson F, Klevmarken A (2007) Comparing register and survey wealth data. Working Paper. Uppsala University, Uppsala

Kalmijn M (2015) Relationships between fathers and adult children: the cumulative effects of divorce and repartnering. J Fam Issues 36(6):737–759

Kantarci T, Van Soest A (2013) Full or partial retirement? Effects of the pension incentives and increasing retirement age in the Netherlands and the United States. Netspar Discussion Paper 10/2013-038. Netspar, Tilburg

Kapteyn A, Banks J, Hamer M et al (2018) What they say and what they do: comparing physical activity across US, UK, and the Netherlands. J Epidemiol Community Health 72(6):471–476

Knoef M, Goudswaard K, Been J et al (2015) Veel variatie in de pensioenopbouw van Nederlandse huishoudens (Heterogeneity in the wealh accumulation of Dutch households). Netspar Brief No. 02. Netspar, Tilburg

Knoef M, Been J, Alessie R et al (2016) Measuring retirement savings adequacy: developing a multi-pillar approach in the Netherlands. J Pension Econ Finance 15(1):55–89

Knoef M, Been J, Caminada C et al (2017) De toereikendheid van pensioenopbouw na de crisis en pensioenhervormingen (Retirement savings adequacy after the crisis and pension reforms). Netspar Design Paper No. 68. Netspar, Tilburg

Kooreman P, Scherpenzeel A (2014) High frequency body mass measurement, feedback, and health behaviors. Econ Hum Biol 14:141–153

Lamers S, Westerhof G, Bohlmeijer E et al (2011) Evaluating the psychometric properties of the Mental Health Continuum-Short Form (MHC-SF). J Clin Psychol 67(1):99–110

Modigliani F, Brumberg R (1954) Utility analysis and the consumption function: an interpretation of cross-section data. In: Kurihara K (ed) Post-Keynesian economics. Rutgers University Press, New Brunswick, pp 388–436

Naumann E, Buss C, Bähr J (2015) How unemployment experience affects support for the welfare state: a real panel approach. Eur Sociol Rev 32(1):81–92

Oude Mulders J, Van Dalen H, Henkens K (2014) How likely are employers to rehire older workers after mandatory retirement? A vignette study among managers. De Economist 162(4):415–431

Scherpenzeel A, Das M (2011) True longitudinal and probability-based internet panels: evidence from the Netherlands. In: Das M, Ester P, Kaczmirek L (eds) Social and behavioral research and the internet: advances in applied methods and research strategies. Taylor & Francis, New York, pp 77–104

Scherpenzeel A, Toepoel V (2012) Recruiting a probability sample for an online panel: effects of contact mode, incentives, and information. Public Opin Q 76(3):470–490

SER (2015) Advies Toekomst Pensioenstelsel (Advice on the future of the Dutch pension system). Social and Economic Council, The Hague

Sonck N, Fernee, H (2013) Using smartphones in survey research: a multifunctional tool. Implementation of a time use app: a feasibility study. Report. Netherlands Institute for Social Research, The Hague

SZW (2016) Perspectiefnota Toekomst pensioenstelsel (Perspective on a new pension system). Ministry of Social Affairs and Employment, The Hague

Trautmann S, Van de Kuilen G, Zeckhauser R (2013) Social class and (un)ethical behavior: a framework, with evidence from a large population sample. Perspect Psychol Sci 8(5):487–497

Van Ingen E, Van der Meer T (2016) Schools or pools of democracy? A longitudinal test of the relation between civic participation and political socialization. Polit Behav 38(1):83–103

Van Ooijen R, De Bresser J, Knoef M (2017) Gezondheid, vermogen en bestedingen van ouderen – beleidsimplicaties voor zorg en pensioen (Health, wealth, and expenditures of the elderly – policy implications for health care and pension). Netspar Brief. Netspar, Tilburg

Marcel Das holds a PhD in Econometrics from Tilburg University, the Netherlands (1998). In 2000, he became the director of CentERdata, a survey research institute specialized in web-based surveys and applied economic research. As a director of CentERdata, he has managed a large number of national and international research projects. He is one of the principal investigators of the Dutch MESS project for which CentERdata received major funding from the Dutch Government. Since February 2009, Das is Professor of Econometrics and Data Collection at the Department of Econometrics and Operations Research of the Tilburg School of Economics and Management at Tilburg University. He has published a number of scientific publications in international peer-reviewed journals in the field of statistical and empirical analysis of survey data and methodological issues in web-based (panel) surveys.

Marike Knoef is Professor Empirical Microeconomics at Leiden University and board member of the Network for Studies on Pensions, Aging and Retirement (Netspar). In addition, she is a fellow at the Research Centre for Education and the Labor Market (ROA). She holds a PhD in Econometrics from Tilburg University. Before Marike joined Leiden University, she worked at CentERdata. Furthermore, she gained experience at the Netherlands Bureau for Economic Policy Analysis and the Dutch Social and Economic Council. Marike's research interests include household saving behavior, economics of aging, labor economics, and health. She gives master classes on these topics at the TIAS School for Business and Society. Recently, she was granted a subsidy for one of her research projects "Uncertainty over the life cycle: implications for pensions and savings behavior."

Part III
Counterfactual Studies

Public Policy, Big Data, and Counterfactual Evaluation: An Illustration from an Employment Activation Programme

Pedro S. Martins

1 Introduction

Joblessness can be a major source of a number of wide-ranging individual and social problems, including poverty, loss of skills, crime, poor health and unhappiness. A key goal of public policy in many countries is therefore to ensure that the labour market operates smoothly and delivers low levels of unemployment over the business cycle. In this context, active labour market policies have been regarded as an important tool that governments can use to increase employment levels. These active policies translate into a potentially large number of programmes, typically led by public employment services, sometimes in partnerships with private providers. Examples of such programmes include registration, profiling and matching of both jobseekers and vacancies, traineeships, hiring subsidies, vocational counselling,

The author acknowledges the financial support from the European Union under the 'ActiValuate: Counterfactual impact evaluation of a large activation programme' project (Directorate-General for Employment, Social Affairs and Inclusion, European Commission), and data access that was provided by the Ministry of the Economy and Employment, Portugal. This paper draws on several results presented in Martins and Pessoa e Costa (2014). Sole responsibility lies with the author: the European Commission is not responsible for any use that may be made of the information contained herein. The author was Secretary of State for Employment in the Government of Portugal between 2011 and 2013 and responsible for designing the programme evaluated here.

P. S. Martins (✉)
School of Business and Management, Queen Mary University of London, London, UK

NovaSBE, Lisbon, Portugal

IZA, Bonn, Germany
e-mail: p.martins@qmul.ac.uk; http://webspace.qmul.ac.uk/pmartins

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_10

training, recognition of prior learning, entrepreneurship support, workfare, and job search monitoring and sanctions (OECD, 2007).

Given the importance of these programmes—Organisation for Economic Co-operation and Development (OECD) member countries spend an average of 0.5% of their gross domestic products (GDPs) on such policies (OECD, 2013)—several evaluations have been conducted across different countries and time periods, in particular from a labour economics perspective (see Martins and Pessoa e Costa (2014) for references). This chapter contributes to this literature by discussing one specific activation programme, but from a broader perspective, emphasising issues of public policy, data and evaluation methods. Instead of focusing exclusively on the econometrics of the evaluation of the programme, as done in most of the literature and in a companion paper (Martins and Pessoa e Costa, 2014), the analysis presented here also draws on the author's role in the design and the implementation of the programme.

The programme evaluated here was based on requiring that certain groups of unemployed individuals in receipt of unemployment benefits participate in meetings in job centres at specific times in their jobless spell. This programme, *Convocatórias* ('summonings'), emerged following from the realisation that the degree of support provided to jobseekers in Portugal was particularly low, with an average of fewer than three job centre meetings per year, compared with averages of more than one per month in some OECD countries. During these new job centre meetings, the jobseekers would be directed towards active labour market measures, including counselling, traineeships, job subsidies, training or workfare. The specific activities would depend on the individual assessment conducted by the jobseekers' caseworkers, including further monitoring of the job search efforts conducted until then, and on the measures available in each job centre. Some unemployed individuals would also be directed towards job interviews, if good matches with available vacancies could be found.

As indicated above, and crucially for identification purposes, the programme was targeted at specific groups of unemployed individuals. These groups were unemployment benefit receivers (UBRs) of a certain age (45 years or above) and those receiving unemployment benefit (UB) for a particular duration (6 months or more). These criteria establish clear differences in programme eligibility across UB duration levels, which are explored in the counterfactual evaluation of the programme through a regression discontinuity (RD) approach (Hahn et al., 2001; Lee and Lemieux, 2010). In particular, the focus is on those aged 44 or below who are targeted exclusively by the UB duration criteria. The effects of the programme, in terms of re-employment and other outcome variables for UBRs unemployed for 6 months or more, in comparison with UBRs employed for less than 6 months, are presented. Given that not all eligible jobseekers actually participated in the programme, owing to capacity constraints, the estimates need to be adjusted in terms of the fuzzy RD approach, as described below.

The empirical analysis draws on two detailed administrative datasets, each including longitudinal individual information on the population of those unemployed over the first 12 months of the programme. The first dataset is drawn from the records of the public employment service and includes information such as the date of registration in the job centre and the date when the unemployed person came into contact with the *Convocatórias* programme (if applicable), as well as several individual variables. The second dataset is extracted from the social security records and includes records on the employment status, salary and UB status of each individual in each month. The two anonymised datasets were merged at the individual level, allowing individuals to be followed from the point at which they became unemployed, through their involvement or not with the *Convocatórias* intervention, and eventually to their return to employment, or not.

The results, presented in detail in Martins and Pessoa e Costa (2014), indicate that the increased activation efforts delivered by the programme had large positive effects in terms of re-employment. This is an important result, particularly given the challenging economic and labour market conditions and the relatively light nature of the intervention. In fact, the estimates imply a doubling of the probability of next-month re-employment for those subject to the programme. The effects estimated are typically of at least 4%, a percentage that exceeds the average monthly re-employment probability over the relevant unemployment duration range.

2 The *Convocatórias* Programme

The background to the programme studied here can be found in an action plan launched by the Portuguese government in March 2012 aimed at the modernisation of the public employment service (PES). This plan included a number of measures, most of which were directed at increasing the activation of the unemployed. The programme studied in this chapter, *Convocatórias*, is one such measure and is based on the requirement that the Portuguese PES (IEFP) calls up all UBRs of specific profiles for meetings with caseworkers in job centres. Moreover, the programme allows job centres to establish the content of meetings and their follow-up, subject to the broad guidelines that the PES should take actions that can activate UBRs and increase their rates of transition to employment. *Convocatórias* strengthened both the extensive and intensive margins of activation, by widening the range of UBRs subject to job centre meetings and by increasing their involvement in active labour market policies (ALMPs), respectively.

In practical terms, the content of the initial meetings and their follow-up actions were varied, depending on the specific profile of each unemployed individual. In general, the job centres monitored the job search effort exerted by the UBR and updated their records regarding the profile of the UBR with a view to facilitating matches with available vacancies. On several occasions, the UBR's personal employment plan, which sets requirements such as a minimum number of monthly job applications to be submitted, was also updated. Moreover, depending on the specific profile of each individual, the job centre would conduct a number of additional actions. These included job search counselling, job interview participation requirements, training, self-employment support, and workfare or traineeship placements.

An additional important aspect concerns the UBR profiles targeted by the programme. Two specific groups were considered, namely UBRs aged 45 or older and UBRs unemployed for at least 6 months. These two groups were considered to be of greater interest in terms of more intense activation work to be delivered by the PES. Moreover, from an operational perspective, the *Convocatórias* programme was implemented gradually, given capacity restrictions across job centres, in some cases also involving a greater priority being given to meetings with UBRs of lower schooling levels. This chapter focuses exclusively on the second group (subsidised unemployment spells of 6 or more months), in particular UBRs with unemployment spells of between 1 and 12 months and not older than 44 years. The latter restriction ensures an exclusive focus on the UBRs subject only to the 6-month stream of *Convocatórias*. The group of unemployed aged 45 or older is more challenging to examine, at least based on the RD approach used here, given that individuals who register as unemployed when they are 45 or older are typically entitled to unemployment benefit for a longer period.

Overall, the *Convocatórias* programme introduced an important strengthening of the activation efforts delivered by the Portuguese PES towards UBRs and the long-term unemployed, involving over 240,000 individuals over its first year of operation.

3 Data

This study draws on two administrative datasets, each one including rich, longitudinal monthly individual information on the population of individuals unemployed at least once over the first 12 months of the programme. The first dataset was drawn from the records of the PES (IEFP) and, in its original version, includes the stock of all individuals registered as unemployed in February 2012 plus the flows of all newly registered unemployed persons from March 2012 up to March 2013. Most activities that were conducted by job centres over that period are also recorded, such as interviews, job placements, training placements and deregistrations, including the specific *Convocatórias* intervention studied in this chapter. The data also include additional information such as the full dates of registration in the job centre and when the unemployed individual was subject to each intervention, as well as several background variables at the individual level, including gender, age, schooling and marital status.

The second dataset was drawn from the records of the social security data agency (II). These data include information on the employment status of each individual in each month over the period under analysis, as well as all earnings, social security contributions and UBs registered. The two datasets were then merged, creating a new dataset that follows individuals as they are unemployed and eventually return to the labour market (such as several of those who were unemployed in February 2012) or are employed, become unemployed and eventually return to employment (such as those individuals who were first unemployed at some point from March 2012).

The merged dataset contains one observation for each individual in each month from February to December 2012. Some individuals or observations are eliminated from this dataset, leaving the final sample that is used to estimate results. First, given that the *Convocatórias* programme was targeted at subsidised unemployed persons, only individuals who have been enrolled in the PES and have received regular UB at least once during the reference period are kept in the sample. Individuals whose potential maximum UB duration is shorter than 12 months are also excluded, because, as UB potential duration influences transitions to employment, they may not be comparable to those who potentially receive UB for a longer period.

As mentioned previously, *Convocatórias* has two eligibility criteria: UBRs who are 45 years or older and UBRs unemployed for at least 6 months. Because the aim is to focus exclusively on those eligible through UB duration, the sample excludes all UBRs who are at least 45 years old (these UBRs would automatically have been eligible as soon as the programme was introduced, implying the need for a different identification strategy). Moreover, the focus here is on UBRs who receive UB for a maximum period that is neither shorter nor much longer than the threshold level of 6 months, given the use of the regression discontinuity approach. Therefore, exclude all observations relating to individuals who received UB for more than 12 months are excluded. Finally, given the focus on transitions out of unemployment, namely of those subject to the programme, the final sample considers only the observations relating to individuals who are unemployed, keeping a record of the timing of a possible transition to employment.

Following the adjustments above, the sample contains 105,595 individuals and 611,061 (individual-month) observations. A total of 25,241 individuals (24%) were subject to the programme. The difference between this figure and the 80,000 reported above is driven by the focus on the stream of the programme targeted at the unemployed on UB for at least 6 months (and aged 44 or below) and the related elimination of those in receipt of UB for more than 12 months. As this programme was targeted initially at unemployed persons, many of whom receive UB for 12 months or more, this naturally leads to a smaller treatment group under analysis. Other individuals are dropped because of data issues, including those who receive different types of UBs (i.e. income support) or exhibit several changes between employment and unemployment over the period considered.

As to the variables used, the main outcome considered is the transition from (subsidised) unemployment to employment, a dummy equal to 1 if a UBR becomes employed in the following month. Other related outcome variables, such as transitions out of subsidised unemployment and transitions to non-subsidised unemployment, are also considered. As in the main case, the transition is assessed from the perspective of the month when the individual is still in a subsidised unemployment situation and analyse possible changes in that situation over a 1-month time window. A fourth dependent variable concerns the income of the individual over the following month. This variable can increase, for instance when a UBR takes a job that pays a salary higher than the UB, or fall, for instance when a UBR moves to non-subsidised unemployment.

The treatment variable is a dummy equal to 1 if the individual was treated, that is, was required to attend a meeting at a job centre under the context of *Convocatórias*, in that month. The analysis also draws on an eligibility variable, which is a dummy indicating whether or not the unemployed person's receives UB for 6 months or more. This variable will be used as an instrument for the treatment, in the context of the fuzzy RD approach (see below). Moreover, several potential explanatory variables are considered: age; gender (a female dummy variable); marital status (married or cohabitant); nationality (foreigner); and schooling years. Other variables used are the potential UB duration and the daily UB amount. These indicate, respectively, the number of days of UB and the amount in euros the unemployed person is entitled to at the time they become unemployed.

Table 1 presents descriptive statistics on all the variables mentioned above on all the observations of the sample used for estimations. It is found that the probability of re-employment in the following month is only 4.4%, while the probability of a transition out of unemployment is 6.2% (and the probability of a transition to non-subsidised unemployment is 1.7%). Average monthly income increase is 1%.

Table 1 Descriptive statistics, pooled data

	Means	S.D.
Transition to employment	0.044	(0.21)
Transition out of unemployment	0.062	(0.24)
Transition to non-subsidised unemployment	0.017	(0.13)
Income percentual variation	0.010	(0.58)
Treatment variable	0.041	(0.20)
Eligibility	0.563	(0.50)
UB elapsed duration	6.21	(2.76)
Age	34.50	(5.66)
Female	0.497	(0.50)
Married	0.514	(0.50)
Foreigner	0.059	(0.24)
Schooling	10.01	(3.86)
Initial UB duration	588.94	(142.10)
UB daily amount	17.57	(6.79)
Observations	611,061	

Source: Martins and Pessoa e Costa (2014). Statistics based on pooled monthly data, from February 2012 to February 2013. Transitions measured in terms of following month. Eligibility is dummy variable equal to one if UB duration is 6 months or more. Schooling measured in years. Initial UB duration denotes maximum number of days of unemployment subsidy at the beginning of the spell. UB daily amount denotes euros per day of unemployment subsidy, at the beginning of the spell

4 The Regression Discontinuity Approach

The analysis of the effects of the *Convocatórias* programme is based on regression discontinuity (RD). Econometric identification in this case draws on the treatment discontinuity that occurs at the UB duration of 6 months. Indeed, the unemployed are eligible only when their spell in receipt of UB hits that threshold.

Before explaining the approach in greater detail, some key concepts and their notation are introduced. First of all, the so-called forcing (or running) variable, Z_{it}, is the UB duration of individual i at month t. This is the variable that will determine the participation in the programme, at some specific value only. (Moreover, to facilitate the interpretation of results, the forcing variable is centred using instead $\widetilde{Z}_{it} = Z_{it} - Z_0$, where Z_0 is the discontinuity point ($Z_0 = 6$, in this case).) The treatment status variable, denoted by D_{it}, is a dummy variable equal to 1 if individual i is called to a *Convocatórias* job centre meeting in month t. Moreover, the outcome variable, denoted by Y_{it}, is a dummy variable equal to 1 when the unemployed individual became employed in month $t + 1$.

As mentioned above, not all eligible individuals are treated at that point. Indeed, as *Convocatórias* was implemented gradually, not every UBR participated in the programme as soon as they became eligible. Hence, the probability of treatment does not jump from 0 to 1 at the specific UB duration threshold, $E_{it} = 1[\widetilde{Z}_{it} \geq 0]$, as in a 'sharp' RD. Instead, the probability increases from zero to a significant positive value at the eligibility threshold—the case of a 'fuzzy' RD design. This jump is illustrated in Fig. 1, which presents the percentage of the unemployed at each UB duration level that are subject to the programme (dots). The figure indicates that the probability of being treated is zero up to the threshold and then jumps to about 0.1 at that level.

The main assumption of the RD approach is that the forcing variable is continuous around the threshold. This assumption is not directly testable, but a graphical analysis is a useful check. Figure 1 also indicates the number of observations for each value of the forcing variable (solid line): unlike in the case of treatment, the evidence is in favour of the continuity of the forcing variable around the threshold.

It is important to note that the profiles of the unemployed persons present will typically be different in each level of unemployment duration, in terms of both observable and unobservable characteristics. This will drive the duration dependence commonly observed in outflows, from some combination of direct effects from unemployment duration (in terms of reduced human capital, for instance) and composition effects (in terms of greater prevalence of individuals who are less likely to find jobs at all levels of unemployment). However, to the extent that the 6-month threshold considered here is not associated with systematic differences across the unemployed in terms of their likelihood of finding employment other than through the effects of the programme (which is indeed the case, to the best of the author's knowledge), the results can be interpreted as the causal impact of *Convocatórias*.

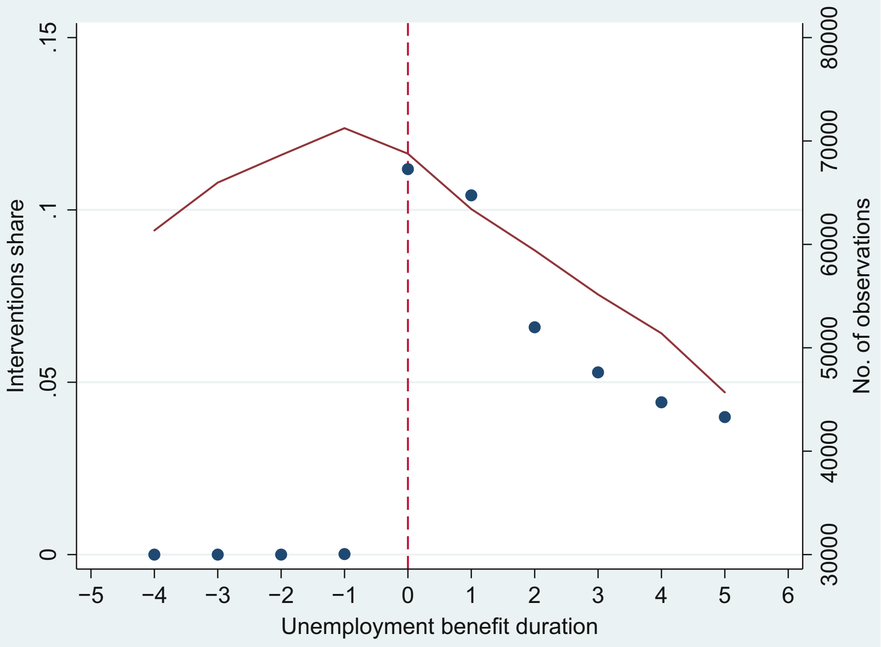


Fig. 1 Probability of treatment and number of observations by (centred) unemployment benefit duration. Source: Martins and Pessoa e Costa (2014). The horizontal axis indicates the (centred) values of UB duration (for instance, zero corresponds to 6 months of UB and 6 corresponds to 12 months of UB). The left vertical axis (and the blue dots) indicate the percentage of observations that are subject to a *Convocatórias* intervention. The right vertical axis (and the red line) indicates the total number of observations used in the pooled cross-section analysis at each specific level of the centred UB duration distribution

More specifically, in the case described here, the fuzzy design is implemented econometrically in terms of two-stage least squares (2SLS), by estimating the following equation:

$$Y_{it} = \alpha + \beta D_{it} + S(\widetilde{Z}_{it}) + \delta X_{it} + \epsilon_{it} \tag{1}$$

in which E_{it} is used as an instrument for D_{it} and X_{it} is a vector of covariates (gender, age, etc.), while $S(Z_{it})$ is a polynomial function of the (centred) forcing variable.

Given that, in the fuzzy design, the probability of being treated is no longer a deterministic function of the forcing variable, the discontinuity in the outcome variable at the threshold cannot be interpreted as an average treatment effect. Nevertheless, Hahn et al. (2001) show that it is possible to recover the treatment effect by dividing the jump in the outcome variable at the threshold by the jump in the probability of treatment, also at the threshold. The latter increase in the probability of treatment is driven by the fraction of individuals induced to be treated ('compliers') who would not be treated in a setting without treatment. This treatment effect is a weighted local average treatment effect (weighted LATE), where the weights are the ex ante likelihood that the individual's Z_{it} is near the threshold.

5 Results

Given the important visual component of a regression discontinuity analysis, this section begins by presenting the graphical evidence of the effects of the *Convocatórias* programme on a number of variables regarding the programme's target group. Specifically, Fig. 2 describes transitions to employment in proportion of the unemployed at different UB duration levels, pooled across different months over the period covered. The figure also includes solid lines on either side of the

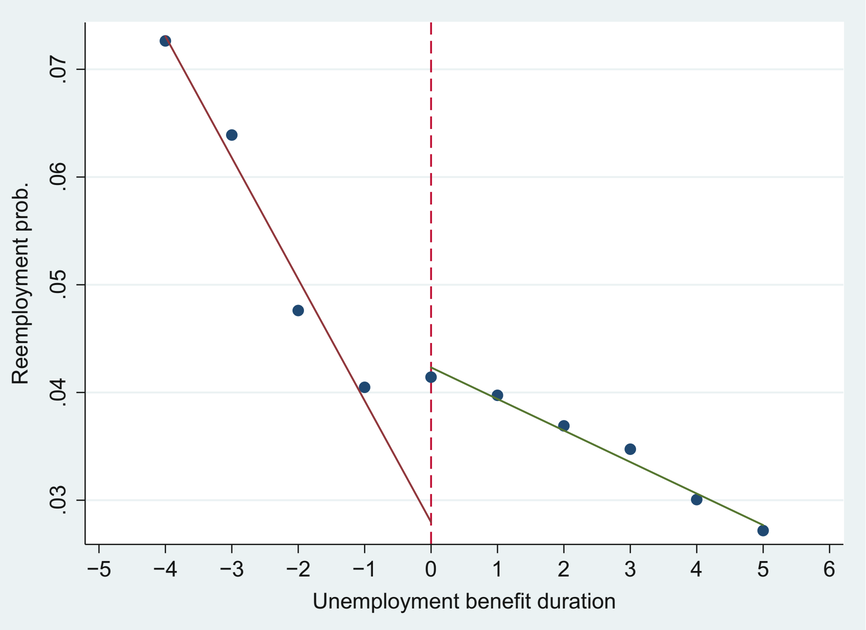


Fig. 2 Re-employment probabilities by (centred) unemployment benefit duration. Source: Martins and Pessoa e Costa (2014). The horizontal axis indicates the (centred) values of UB duration. The vertical axis indicates the probability of re-employment in the subsequent month. The red and green lines correspond to fitted linear equations over the four and five observations at the left and right of threshold UB duration, respectively. The left line was extended towards the threshold value by computing its predicted value at that level of UB duration

threshold obtained from linear splines estimated, respectively, over the (centred) UB duration intervals [-4;0] and [0;5].[1]

Figure 2 presents a downward trend in re-employment probabilities (average transitions to employment) as UB duration increases, consistent with evidence of negative duration dependence observed in studies of unemployment. However, at the threshold UB duration, there is graphical evidence of a (discontinuous) increase in the re-employment probability, after which it resumes its downward trend, although at a flatter rate. Moreover, the gap between the predicted re-employment probability and the actual value at the threshold unemployment duration is sizable, about 1 percentage point. In the context of the RD approach, this discontinuous increase in the re-employment probability can be interpreted as a treatment effect, especially after adjusting for the fact that many eligible individuals were not treated. A very similar pattern in transitions out of unemployment is seen. The average values are higher across the range of unemployment durations given the wider coverage of outcomes (transitions to both employment and non-employment), ranging from 5 to 8%, while before this range was 3 to 7%, but again there is a pronounced discontinuity at the threshold duration.

On the other hand, it is observed that transitions to non-subsidised unemployment increase steadily with UB duration, although at a lower probability than transitions to employment. More importantly, there is virtually no discontinuity at the 6-month threshold, nor any sizeable change in the slopes of the best-fit lines. Finally, there is no evidence of a discontinuous change in income levels at the threshold UB duration.

The robustness of the graphical evidence above is now tested, estimating the model described in Sect. 4. In particular, 2SLS models with a linear spline, $S(\widetilde{Z}_{it}) = \pi_0 \widetilde{Z}_{it} + \pi_1 \widetilde{Z}_{it} D_{it}$, are estimated, using the eligibility variable E_{it} as an instrument for the treatment variable D_{it} resulting in the following second-stage equation:

$$Y_{it} = \alpha + \beta \hat{D}_{it} + \varphi_0 \widetilde{Z}_{it} + \varphi_1 \widetilde{Z}_{it} \hat{D}_{it} + \delta X_{it} + \epsilon_{it} \tag{2}$$

The key dependent variables considered, Y_{it}, are, in turn, the transitions to employment (re-employment probability), to being out of unemployment, to non-subsidised unemployment, and the income variation in the following period. The remaining terms of the equation are the same as explained above. The coefficients on the treatment effects are the β's for each equation, according to the outcome variable.

The results following the estimation of the model above are presented in Table 2. Also presented are the results from different spline specifications (across rows). The first-stage estimates are presented in the last column and are the same for all outcome variables. Each coefficient and standard error pair across the first four columns corresponds to a separate estimation of a different model in terms of

[1] See Martins and Pessoa e Costa (2014) for additional figures on transitions out of unemployment, transitions to non-subsidised unemployment and income level percentage changes.

Table 2 *Convocatórias* programme effects, different dependent variables and polynomials

Polynomial function	Treatment effect on Re-employment probability	Transitions... Out of unemployment	Transitions... To non-subsidy unemployment	Income level	1st stage results—eligibility effect
Linear	0.021*** (0.007)	0.023*** (0.008)	0.003 (0.004)	0.007 (0.022)	0.130*** (0.001)
Quadratic	0.083*** (0.009)	0.080*** (0.010)	−0.003 (0.004)	−0.011 (0.026)	0.114*** (0.001)
Cubic	0.040*** (0.009)	0.047*** (0.010)	0.008** (0.004)	0.017 (0.028)	0.133*** (0.001)
Linear spline	0.041*** (0.009)	0.041*** (0.009)	0.000 (0.004)	0.002 (0.023)	0.106*** (0.001)
Quadratic spline	0.097** (0.040)	0.099** (0.041)	0.002 (0.012)	0.237** (0.109)	0.110*** (0.001)
Outcome mean	0.044	0.062	0.017	0.010	–
Obs.	611,061	611,061	611,061	600,412	–

Source: Martins and Pessoa e Costa (2014). Each coefficient and standard error pair is obtained from a separate 2SLS regression under a specific spline structure (indicated in the left column) and dependent variable (indicated in the top row). The last column presents the results for the first-stage results on programme eligibility term without interactions, under each polynomial function. All specifications include a large set of control variables (see main text). Standard errors in parentheses $^{**}p \leq 0.05$; $^{***}p \leq 0.01$

the outcome variable and the polynomial function. Turning to the analysis of the estimates, considering the first column, which focuses on the key dependent variable (transitions to employment), the results across polynomials confirm the graphical evidence in Fig. 2 and support its robustness. In all models, participation in the *Convocatórias* programme results in significantly positive effects in terms of re-employment probabilities. The magnitude of the coefficients varies from 2% (linear polynomial) to 9% (quadratic spline). These coefficients represent an increase in re-employment probabilities from 50 to 225%, taking into account the outcome mean of 4.4% (see the second last row in the table).

In terms of the remaining dependent variables, the results on the transitions out of unemployment are very similar to those on the equivalent specification for transitions to employment, as predicted from the graphical evidence. In the case of transitions out of unemployment, the coefficients also range between 2 and 9%. Consistently, the transitions to non-subsidised unemployment are found not to be affected by the programme, with virtually all results insignificant. In other words, most individuals who leave unemployment find jobs. Similarly, no effects are found in terms of income variation, defined as the percentage change in the sum of all UB and employment earnings. The last result indicates that employment earnings obtained are similar to the income wage from benefits, which is consistent with the generally high levels of replacement rates (nearly 100% for individuals on low wages) and the previous results about most transitions being to employment (rather than to non-employment).

It is also important to note that the first-stage coefficients on eligibility (the instrument) are always significantly positive, with coefficients of around 12%, and little variability across polynomial functions. This result confirms the relevance of the eligibility status as established in the programme in terms of actual participation in *Convocatórias*, namely through a request that the UBR attends a job centre meeting.

Overall, the findings on re-employment effects can be regarded as larger than those commonly found in the literature. Of the 15 studies surveyed, only Geerdsen (2006) has a similar magnitude, although that programme was implemented when the unemployment rate was only 6.1%. One explanation may be related to the relatively light activation efforts that had been conducted, in general, by the Portuguese PES up until the introduction of *Convocatórias*, especially following the large increase in the number of unemployed and the decline in vacancies. This situation may give rise to higher than average marginal re-employment benefits even from relatively moderate levels of activation, such as interviews for available vacancies or 1-day job search training sessions, despite the poor labour market conditions.

Another related explanation concerns the role of 'threat' effects (Black et al., 2003). As the *Convocatórias* programme consists of a meeting with a caseworker, generally followed by referrals to ALMPs, some UBRs may perceive participation as an increased cost of being unemployed. Those UBRs may therefore increase job searches and/or decrease their reservation wage even before participation or soon after it begins, leading to the documented increase in transitions out of unemployment. Moreover, the programme may have prompted some targeted UBRs who were employed informally to stop collecting UB and to register their jobs with social security instead, given the impending likelihood that they would be required to participate in training or workfare, for instance. On the other hand, as mentioned above, the results on transitions out of unemployment are exclusively driven by an increase in re-employment probabilities and not by an increase in transitions to non-subsidised unemployment, unlike in Manning (2009) and Petrongolo (2009).

6 Lessons Drawn

This chapter illustrates the strong interplay between policy, data and (counterfactual) evaluation with an illustration from a key social and economic area—the labour market. It traces the development, implementation and evaluation of an ALMP in Portugal, which involved the participation of over 200,000 individuals in its first year of operation alone.

One first important lesson that can be drawn from this process concerns the value of international benchmarking and benchlearning exercises. In the case of the *Convocatórias* programme, its motivation largely stemmed from the awareness that there were important gaps in the support provided to jobseekers in the country compared with provision in other OECD economies. In fact, before the programme, jobseekers in Portugal had very few meetings with representatives of the PES, and very little job search monitoring took place. This contrasts with the roles of public employment services in a number of other EU countries, especially those that may be regarded as closer to the technological frontier in this area. A considerable part of the success of this policy can be attributed to the gap between the intensity of previous activation practices and the intensity of practices implemented following the deployment of the programme, something which would not be apparent without a benchmarking exercise.

To take this point one step further, additional efforts towards clearer benchmarking and benchlearning in the many specific dimensions of the work conducted by public employment services, ideally involving greater interaction with research centres and their analytical perspectives, may pave the way for further improvements in the EU and elsewhere in terms of the support provided to jobseekers and in pursuit of lower unemployment. Such benchmarking exercises are of course a key area from the perspective of data, even if not necessarily in terms of counterfactual evaluation methods.

A second lesson concerns the importance of building in an evaluation component in new public policies even before they are implemented. This perspective can greatly facilitate a rigorous evaluation, in contrast to attempts at measuring impacts on a strictly ex post basis. In the case of *Convocatórias*, while the sharp discontinuities at the 6-month unemployment duration threshold were introduced to facilitate a counterfactual approach, in hindsight more could have been done to enhance the insight provided by the programme and prevent some pitfalls, including the mismatch between job centre capacity and eligible jobseekers. For instance, a more staggered approach towards the roll-out of the programme, because of the capacity issues above, could have been exploited in terms of greater randomisation. In other words, given that not all jobseekers could have been supported immediately, randomisation of targeted jobseeker profiles across the 80 job centres would have greatly increased the potential insight from the evaluation, for instance in terms of the potential interactions between profiles and impacts.

A related lesson is about the effects of the evaluation of a programme on the impact of the programme itself. Although this ’impact of impact’ parameter cannot easily be measured quantitatively, it might be argued that, when the main stakeholders involved in the implementation of a programme are aware of its ongoing rigorous evaluation, there are stronger incentives for a more successful implementation of that same programme. On the contrary, when no such rigorous evaluation is in place, the impact of the programme may suffer. This may also generate a form of ‘publication bias’, whereby evaluated programmes will tend to add greater value than non-evaluated programmes.

Finally, a fourth lesson from this case study relates to the critical relevance of quality microdata, perhaps ideally based on administrative sources. The evaluation conducted here was based on the matching of two individual-level longitudinal datasets, one collected by social security, the other by the public employment services. In both cases, these data would be collected with or without the programme and therefore did not generate opportunity costs, other than the anonymisation and matching processes. On the other hand, their analysis, through state-of-the-art, transparent regression discontinuity methods, offered great insight in terms of the impact of the underlying programme. Moreover, the public good nature of such (big) data—as they are 'non-rival in consumption'—also enhances their economic impact: many individuals can use them without additional costs of production. Additional efforts by public agencies towards making such anonymised datasets freely available can become an important source of value added, economic growth and also higher levels of employment.

References

Black D, Smith J, Berger M, Noel B (2003) Is the threat of reemployment services more effective than the services themselves? Am Econ Rev 93(4):1313–1327

Geerdsen L (2006) Is there a threat effect of labour market programmes? A study of ALMP in the Danish UI system. Econ J 116(513):738–750

Hahn J, Todd P, Van der Klaauw W (2001) Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69(1):201–209

Lee D, Lemieux T (2010) Regression discontinuity designs in economics. J Econ Lit 48(2):281–355

Manning A (2009) You can't always get what you want: the impact of the UK Jobseeker's allowance. Labour Econ 16(3):239–250

Martins PS, Pessoa e Costa S (2014) Reemployment and substitution effects from increased activation: evidence from times of crisis. Discussion Paper no. 8600, IZA, Bonn

OECD (2007) Activating the unemployed: what countries do. In: OECD employment outlook. OECD, Washington, pp 207–242

OECD (2013) Activating jobseekers: lessons from seven OECD countries. OECD Employment Outlook. Organisation for Economic Cooperation and Development, Paris

Petrongolo B (2009) The long-term effects of job search requirements: evidence from the UK JSA reform. J Public Econ 93(11–12):1234–1253

Pedro S. Martins has been a Professor of Applied Economics at Queen Mary University of London since 2009. Research fellow at IZA, Bonn, and NovaSBE, Lisbon. Secretary of State for Employment in the Government of Portugal between 2011 and 2013, responsible for a number of reforms in employment law, collective bargaining, public employment services, and the European

Social Fund. Member of the group of experts advising the Government of Greece and the European Commission on labor market reforms in 2016. Author of over 20 peer-reviewed journal articles on labor economics, economics of education, and international economics.

The Use of Administrative Data to Evaluate the Impact of Active Labor Market Policies: The Case of the Italian *Liste di Mobilità*

Enrico Rettore and Ugo Trivellato

1 Introduction

High-quality data are essential to design an effective evaluation of the effects of a public intervention. Leaving aside the notion of quality relevant in all scientific fields—available information should provide valid and reliable measurements for the concepts they are intended to measure—this chapter takes "high quality" to have two different but equally relevant meanings. First, the data should provide information on the *outcomes* relevant for the evaluation of the intervention, i.e., the individual status and behaviors the intervention might have an impact on, whether intentionally or as a side effect. Second, the data should allow identification of a *comparison group*, made up of individuals *not exposed* to the intervention under evaluation and *equivalent* to the group of individuals exposed to the intervention, in all respects relevant to the outcomes considered in the evaluation.

The first condition—which may seem obvious—often requires a great deal of effort by the analyst to recover information on the outcomes the intervention might have an impact on.

The second condition is more technical. To identify the average causal effect of an intervention on the pool of individuals exposed to it, one needs to properly approximate what those individuals would have experienced in the so-called *counterfactual* world, i.e., one in which the intervention does not happen. This counterfactual experience being by definition unobservable, one has to resort to

E. Rettore (✉)
Department of Sociology and Social Research, University of Trento and FBK-IRVAPP, Trento, Italy
e-mail: enrico.rettore@unitn.it

U. Trivellato
Department of Statistical Sciences, University of Padova and FBK-IRVAPP, Padova, Italy
e-mail: ugo.trivellato@unipd.it

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_11

the *factual* experience of a pool of individuals not exposed to the intervention, the comparison group. To be a credible approximation of the counterfactual, the comparison group must be made up of individuals as similar as possible to those exposed to the intervention.

By far the most convincing way of obtaining two groups that are equivalent in all respects—one exposed to the intervention, the other providing the approximate counterfactual—is to select them at random in the same way as experiments run in laboratories to test the effectiveness of a new drug. In an observational study—i.e., where the analyst has no control over who is going to receive the exposure—an effective impact evaluation design requires the analyst to understand the selection process as well as possible, i.e., the set of rules and behaviors according to which the exposure state of the individuals is determined. Knowledge of the selection process *implies* knowledge of the differences in composition of the two groups of individuals.

This raises an important issue. Knowledge that, as a result of the selection process, individuals exposed to the intervention are on average, for example, younger, more educated, or less experienced, specifies the information required to select a proper comparison group. This entails observing age, education, and previous labor market experience for each individual included in the study. With this information at hand, the analyst is in a position to compare the two groups of individuals, controlling for the characteristics with respect to which the two groups are on average different as a result of the selection process.

Administrative data are a valuable source of information in several ways for the design of an impact evaluation. Their obvious main limitation is that they are developed and maintained to meet the specific administrative requirements of their reference institution and are not primarily intended to serve scientific purposes. As a straightforward implication, the (sub)population covered by the archive and the information it provides may be useless for the specific problem in which the analyst is interested. On the other hand, conditional on providing the information required for the evaluation, administrative data feature major advantages: they cover the whole (sub)population of reference for their administrative purposes, and not just a sample of individuals, very often they extend over long time periods, they are not susceptible to the typical measurement errors of survey data, and, finally, they are free of charge.

To illustrate the potential as well as the limitations of administrative data for evaluating public interventions, this chapter presents the case of the Italian *Liste di Mobilità* (LM), a program to handle collective redundancies which was introduced in the early 1990s and was in operation until recently. All the existing studies on the LM have exploited administrative data, with a clear distinction between first-generation studies, which relied on poor information, and subsequent studies that over the years have had access to much richer data, which gave insight into the effects of the LM unknown to first-generation studies.

This chapter is organized in five sections. Section 2 presents the basic provisions of the LM program. Section 3 describes the evaluation questions on which existing studies focused. Section 4 presents a review of the evaluation designs adopted by researchers and of the results obtained. The emphasis in this section is on the advantages and limitations of the administrative archives from which researchers drew their data. Section 5 concludes. For those unfamiliar with the econometrics of program evaluation, an appendix briefly describes the basic ideas on which the evaluation of LM has so far been developed.

2 Basic Provisions of the *Liste di Mobilità* Program

The LM program was introduced by Law 233 (1991) and then slightly modified by Law 236 (1993). It was in operation until the end of 2016. Some provisions varied according to industry, worker's occupation, geographic area, etc. and underwent frequent modifications over time. Here, the main provisions relevant to the area and the period covered in this chapter are outlined. For additional details, see Paggiaro et al. (2009) and references therein. Firms with more than 15 employees—referred to here as "large" firms—could collectively dismiss redundant workers, because of plant closure or restructuring, and automatically enroll them in a special register maintained by a regional authority. Workers dismissed by firms with up to 15 employees—referred to here as "small" firms—could also enroll in the LM on a voluntary basis. Evidence indicates that most eligible workers dismissed by small firms did register in the LM. To be eligible for the LM, a worker must have had been on a permanent contract with the dismissing firm for at least 1 year. Workers in the LM were in principle required to fulfill some obligations with respect to training and job offers. An LM worker who refused an appropriate job offer from the local public labor exchange was dropped from the program. However, enforcement of these rules was largely absent. In practice, a worker's willingness to accept a job offer was not tested, and a worker enrolled in the LM could refuse any job offer and retain LM status up to the end of the eligibility period. The basic features of the program are summarized in Table 1. It is apparent that there were two separate subprograms, targeted at two non-overlapping populations:

1. The first subprogram applied to workers who had been collectively dismissed by large firms. Upon dismissal, they entered the LM by default and received a monetary benefit, which was partly transferred to any firm that later hired them. The same firm also benefited from a rebate on social security contributions (SSCs) for up to 2 years. This subprogram was fundamentally different for workers aged 50 years or over who met the requirements for "long mobility." For these workers, the active component of the program was largely dominated by the passive component.
2. The second subprogram applied to workers either collectively or individually dismissed by small firms. They entered the LM on a voluntary basis and were eligible only for the active component of the program, which could be supplemented with the much less generous standard unemployment benefits where applicable.

Table 1 Basic features of the Liste di Mobilità (LM)

	Firm size					
	Workers collectively dismissed by large firms (>15 employees)			Workers collectively dismissed by small firms (≤15 employees)		
Age at dismissal	<40	40–49	≥50	<40	40–49	≥50
Eligibility duration (years)	1	2	3	1	2	3
Monetary benefits (replacement rate, with a ceiling)	80%	First year: 80% Second year: 64%		–[b]	–[b]	–[b]
Rebate on SSCs	97% of the standard SSCs, for 18–24 months					
Benefit transfer to the hiring firm	50% of 1 year at most	50% of 2 years at most		–	–	–

[a] Workers ≥50 years old eligible for monetary benefits maintain their eligibility status even longer if they are close to being eligible for retirement benefits (the so-called long mobility)

[b] SSC Social security contribution. These workers may draw the standard UI (over the period of this study, replacement rate 30%), provided they meet the eligibility criteria

Source: Paggiaro et al. (2009)

The common feature of the two subprograms was the duration of eligibility as determined by the worker's age at dismissal. However, the effect of the eligibility duration was very likely to vary across subprograms, as well as across the 40- and 50-year thresholds. Calculations by Paggiaro et al. (2009) show that the best strategy for the employer was to hire a worker from the LM on a temporary 1-year contract and then to switch it to a permanent one. If the worker was not eligible for monetary benefits, this strategy provided a saving over 2 years, worth approximately 23% of the labor cost. The employer saved an additional 13% of the labor cost by hiring a worker eligible for monetary benefits on his or her first day in the LM, irrespective of the duration of eligibility.

The bulk of savings for the hiring firm was represented by the significant rebate on SSCs. In the case of the best hiring strategy, it ranged from 65% to 70% of total savings. In addition, it coincided with total savings—which at 23% of the 2-year labor cost were still substantial—in all circumstances when there was no benefit transfer and specifically at the end of the eligibility period. One implication is that it might not have been all that relevant to a potential employer whether or not an LM worker to be hired was entitled to monetary benefits and whether he or she was above or below the 40-year threshold, because in all cases the employer received the same rebate on SSCs, as long as the worker was eligible.

3 The Evaluation Problem

Most of the studies on the LM so far have focused on the effect of the length of the eligibility period. As an example, looking at the effect of being eligible for 2 years in the LM instead of just 1 year, the issue can be formulated as follows: does allowing workers just above the 40-year threshold to stay in the program for 2 years affect their chance of reemployment—and the quality of the job they eventually find—relative to what would happen to them with 1-year eligibility? With the standard job search model as a background (Mortensen and Pissarides 1999), it is apparent that the LM provides two contrasting incentives: (i) the incentive to firms, which benefit from the rebate on SSCs (and in some cases from the benefit transfer), to provide workers enrolled in the LM with more job offers than they would otherwise receive and (ii) the incentive to workers drawing monetary benefits from the LM program to lengthen their unemployment spell by increasing their reservation wage and to refuse any job offers they receive, at least over a large fraction of their eligibility period.

In comparative terms, both these incentives tend to be higher under the 2-year regime than under the 1-year regime. From a theoretical point of view, indication of the effect is a priori uncertain and depends on which of the two incentives prevails. Furthermore, it is difficult to isolate the effects of these incentives from those of provisions about engagement in temporary employment. Thus, the effect can be identified only empirically.

Research on the LM so far has focused largely on program participants aged under 50 years, focusing on the causal effect of being eligible for 2 years versus 1 year either on the duration of the unemployment spell or on the probability of having working status at selected post-enrollment periods (see Brunello and Miniaci 1997, Paggiaro and Trivellato 2002, Paggiaro et al. 2005). The prevailing evidence can be summarized in two statements: (i) for workers entitled only to the active component of the LM, i.e., dismissed by small firms, the additional year of eligibility has no effect on reemployment probabilities or on the time spent waiting for the first permanent job, and (ii) for workers also entitled to monetary benefits, i.e., dismissed by large firms, the additional year of eligibility has a negative impact: older workers, who draw benefits longer, have significantly lower reemployment probabilities and a significantly longer time spent waiting for a first permanent job than their younger colleagues. This effect tends to be larger for women, but not consistently across all the case studies.

4 How the Availability of Data Drove the Design of the Evaluations

There are three main sources of administrative information relevant to the evaluation of the impact of the LM:

1. The archive resulting from the management of the LM itself
2. The archive resulting from the operations of the public labor exchange (*Centri per l'Impiego* (CPI))
3. The archive of the Italian social security agency (INPS)

These three sources differ greatly in terms of the information they provide on the labor market history of workers enrolled in the LM. Broadly speaking, the first archive provides information only on transitions to permanent employment after enrollment in the LM; the CPI archive is much more detailed, providing information on each single employment spell, whether temporary or permanent, and both before and after enrollment in the LM; and, finally, the INPS archive adds data on employee wages to the CPI information.

The first-generation studies on the LM made use of the first archive only. More recent studies had access to the CPI archives and in some instances to the INPS archive. All these studies are broadly similar in terms of the basic feature of the evaluation design: one way or another, they exploit the discontinuity along the age dimension in the duration of eligibility. But they are quite different with respect to the variety of outcomes taken into consideration as well as with respect to the robustness of the results obtained. All these differences across studies are driven entirely by the administrative information available to the researchers. The main studies are reviewed below, focusing specifically on the value of each piece of information to the design of the impact evaluation.

4.1 Studies Based on the Liste di Mobilità Archive

The LM was managed by regional employment agencies, which were also responsible for data collection. There was no common format for collecting individual data on the program across the country; therefore, there is no consistent national database available in Italy resulting from the operation of those agencies. To illustrate what first-generation studies did, the work of Paggiaro and Trivellato (2002) is considered here. They use data from the administrative records of the Veneto Regional Employment Agency. The Veneto region is a large, relatively well-developed region of Northeastern Italy. With more than 4.4 million inhabitants, it makes up 7.7% of the Italian population. At the time the authors refer to (the late 1990s), Veneto had an employment rate close to 42%, an unemployment rate around 5.2%, and a per capita gross national product (GNP) some 15–20% higher than the national average. These traits of comparatively low unemployment and high economic activity characterize Veneto as similar to the rest of Northern Italy but far from representative of the much less developed South.

The analysis by Paggiaro and Trivellato (2002) is restricted to the period from January 1995 to March 1999. Each worker is followed from enrollment in the LM up to the occurrence of one of the following events: (i) exit into a permanent job and (ii) exit from the lists at expiration of the eligibility period. Incomplete durations are registered for workers still enrolled in the lists at the end of the observation window. There are clear limitations arising from these data.

The main limitation is that there is no explicit information on spells of temporary work. It is clear that some of these spells did exist because there were cases where the duration of enrollment in the lists was longer than the legal duration. This prevented the authors from discriminating between periods spent in the lists as temporarily employed and time spent unemployed (drawing income support, if so entitled). Those periods are then collapsed into a single spell of enrollment in the lists, waiting for one of the two alternative final events, i.e., transition to a permanent job or expiration of the eligibility period.

In principle, there is also a problem of self-selection into the program, since workers dismissed by small firms who decide not to register are not included in the study. Fortunately, in practice this problem turned out to be irrelevant: evidence provided by officials of the Veneto Regional Employment Agency indicates that, essentially, all workers dismissed by small firms did register in the lists.

The analysis focuses on the first two age groups, while workers aged 50 or over are excluded. The final sample is thus reduced to 36,405 workers. The parameter of interest is the effect of being offered 2 years of eligibility instead of 1 on the rate of transition to permanent employment. The analysis is performed separately for workers entitled and not entitled to income support, and by gender. Since duration of eligibility depends only on age at dismissal, the assignment process fits the sharp regression discontinuity design (RDD). In principle, one could exploit this feature to identify and non-parametrically estimate the causal effect of the additional year of eligibility by comparing workers just above the 40-year threshold with those just below it.

Considering the size of the sample and the relatively low rate of transition to permanent employment, the authors chose a semiparametric strategy exploiting the whole sample. They fit a proportional hazard model to the duration of the spell in the lists, estimating the difference between the two groups of workers—up to 39 and 40–49, respectively—controlling for age (as well as for other observable characteristics of the workers). Figure 1 reports their main results.

For workers entitled to income support, the additional year of eligibility appears to have a significant positive effect on the time spent searching for permanent employment, varying with the duration of the spell and by gender. For men, the effect is negligible during the first year but has a strong negative peak just at the end of the first year, i.e., when the eligibility period for younger workers expires. Within the second year, the effect is still negative but less pronounced, possibly because the reference group of workers under 40 is somewhat selected, as it consists only of workers who obtained temporary contracts. The pattern of the effect is quite different for women. It is negative from the start of the spell, stays essentially constant throughout the first year, including the last week, and then increases during the second year up to its end. Similar differential treatment effects, but much less pronounced, are also found for workers dismissed by small firms, for whom benefit packages do not include any benefit transfer component.

Summing up, the passive component of LM appears to drive the results. A major open question is that with these data, there is no way of understanding what happens to the overall employment chances—rather than simply permanent employment—of workers enrolled in the LM, both during enrollment and in the years after expiry of the eligibility period.

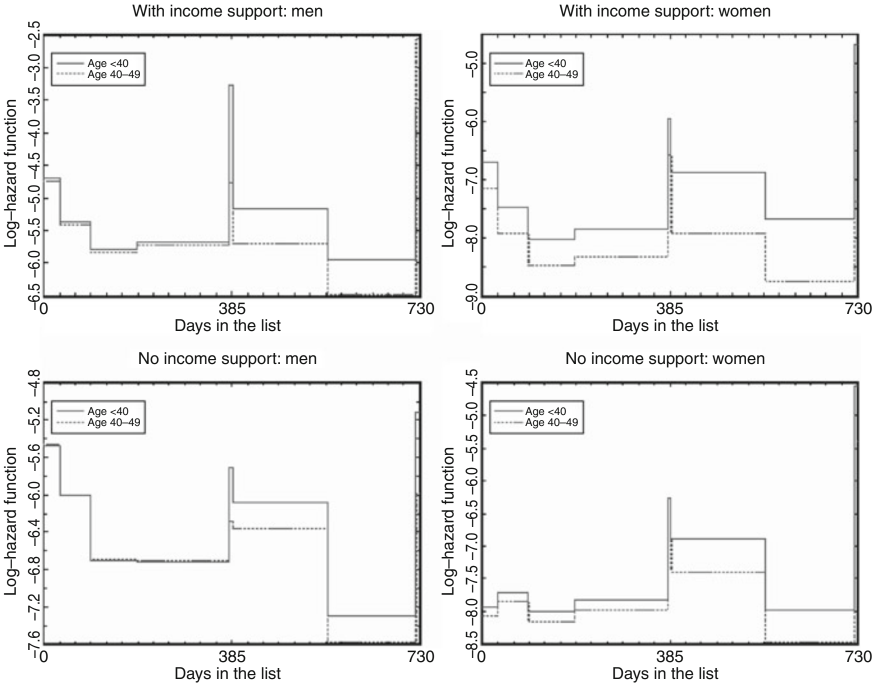


Fig. 1 Estimated baseline hazards controlling for age. Source: Paggiaro and Trivellato (2002)

4.2 *Studies Based on the Archives of the Public Labor Exchange and the Social Security Agency*

Compared with the LM archive, the key additional piece of information provided by the CPI archive is a much more detailed description of the labor market history of workers enrolled in the LM, before, during, and after the eligibility period. Two major advances are thus possible with respect to the first-generation studies. First, the causal effects of the duration of eligibility are established with respect to the *monthly employment rate* in each month since enrollment in the lists. This provides a much richer picture of the short- and medium-term effect of the program on the pattern of reemployment of these workers.

Second, the rich set of information on preenrollment labor market history of workers entering the LM allows implementation of a set of specification tests to validate the design of the impact. This strengthens the credibility of the results.

Finally, access to the INPS archive allows addition of information on wages to the analysis. The next part of this chapter presents the analysis developed by Paggiaro et al. (2009). They refer to workers enrolled in the LM in Veneto, 1995–1998. Their analysis is carried out separately (i) for the two LM subprograms, (ii) by gender, and finally (iii) both at the 40-year and at the 50-year thresholds.

As in the first-generation studies, Paggiaro et al. (2009) identify the causal effect by exploiting the discontinuity along the age dimension in the duration of the eligibility. The internal validity of this strategy rests on the assumption that, if individuals just above and just below the threshold were assigned to the same eligibility regime, they would experience the same average outcome. Clearly, any evidence providing support to this claim would be most welcome. The authors also argue that, at least in principle, one could imagine reasons why the assumption might be violated. For instance, one could argue that workers enrolled in the LM and over the threshold are on average different from those below it, if firms and unions bargain on the composition of the pool of workers to be dismissed.

Figure 2 shows the distribution of enrolled workers by age, separately for the two groups of entitlement to monetary benefits and for gender. The main evidence is a large discontinuity at the 50-year threshold for workers with monetary benefits.

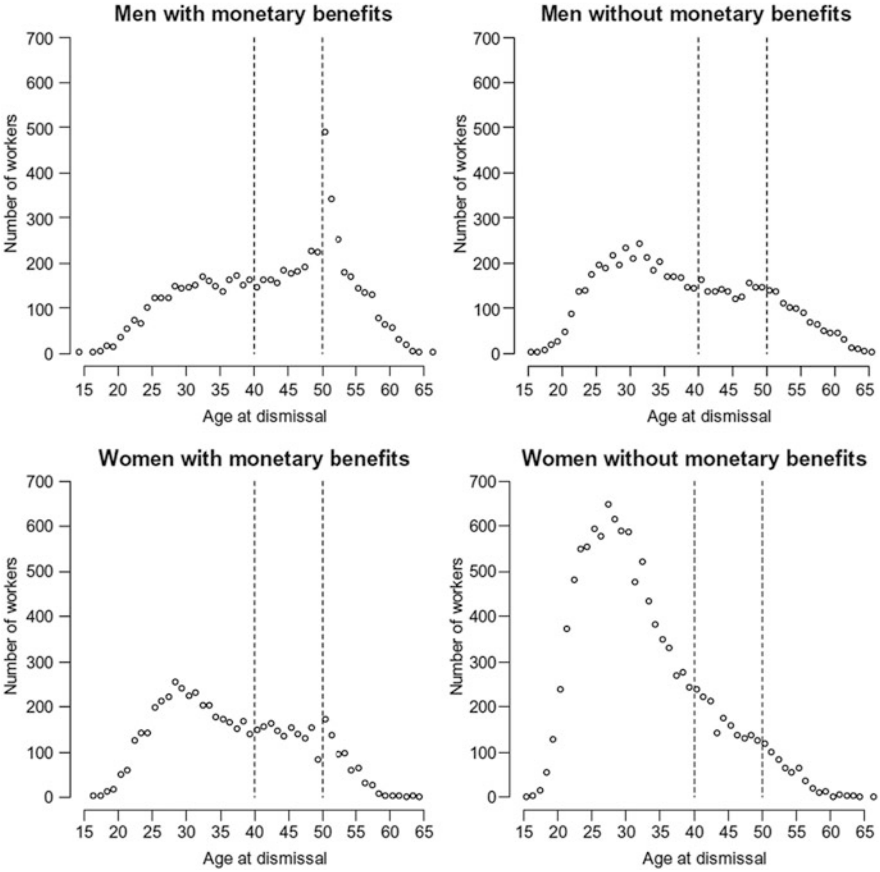


Fig. 2 Distribution of workers enrolled in the *Liste di Mobilità* by age, 1995–1998, Veneto. Source: Paggiaro et al. (2009)

The McCrary (2008) test turns out to be statistically significant for both men and women. As this is the threshold at which most of the workers are entitled to "long mobility," it is clear that workers dismissed by large firms have been selected in consideration of the peculiar differential advantages brought to them by the "long mobility" provisions. In principle, this selection opens the door to a violation of the identifying restriction of the RDD.

To investigate this, the authors follow Lee (2008) and carry out a set of overidentification tests to validate the strategy. These tests are based on comparing individuals just above the threshold and their younger colleagues just below it with respect to their preprogram employment history. As it is hard to think of a causal effect of the LM program on employment status and wages experienced by a worker 3 years *before* entering the LM, for example, any discontinuity with respect to those variables at the cutoff points should be interpreted as a sign of a differential composition around the cutoff point with respect to characteristics relevant to subsequent employment status and wages. Hence, it should be taken as evidence against the validity of the RDD restriction.

Figure 3 plots the employment rate by age for workers enrolled in the LM 3 years before enrollment. No significant discontinuity appears at the cutoff points. Similar evidence is presented in the paper for the weekly wages of workers, as well as looking at other preenrollment periods. This evidence confirms that workers aged 39 and 40, as well as workers aged 49 and 50, had the same employment histories before entering the program.

Figure 4 shows the age profile of the reemployment probability 3 years after enrollment, when most workers have already dropped out of the LM. The age profile shows no significant discontinuity at the thresholds, except for a major statistically significant drop at 50 years for workers with monetary benefits, i.e., those who can use "long mobility" as a bridge to retirement. The drop is as large as 32.6% for men, whereas it is 27.1% for women.

Having established that adding a second year of eligibility did not affect the chances of getting a job, the issue is whether or not it affected the quality of the jobs these workers in their early 40s eventually found. In particular, did it have any effect on wages once a worker enrolled in the LM secured a new job? Keeping the standard job search model as a background, it might well be the case that entitlement to a longer eligibility period allows a worker to be more selective during the job search, particularly if he or she draws monetary benefits, ultimately obtaining a higher wage.[1]

The evidence presented by Paggiaro et al. (2009) shows that there are no significant effects for all the groups taken into consideration. The only slight difference discernable is again at the 50-year threshold for both men and women with monetary benefits. However, this difference is hardly significant and is of the

[1] This issue raises an additional econometric problem pointed out by Ham and Lalonde (1996). The authors argue that the problem does not arise in this specific case, since there is no causal effect (at least on average) on the probability of having working status.

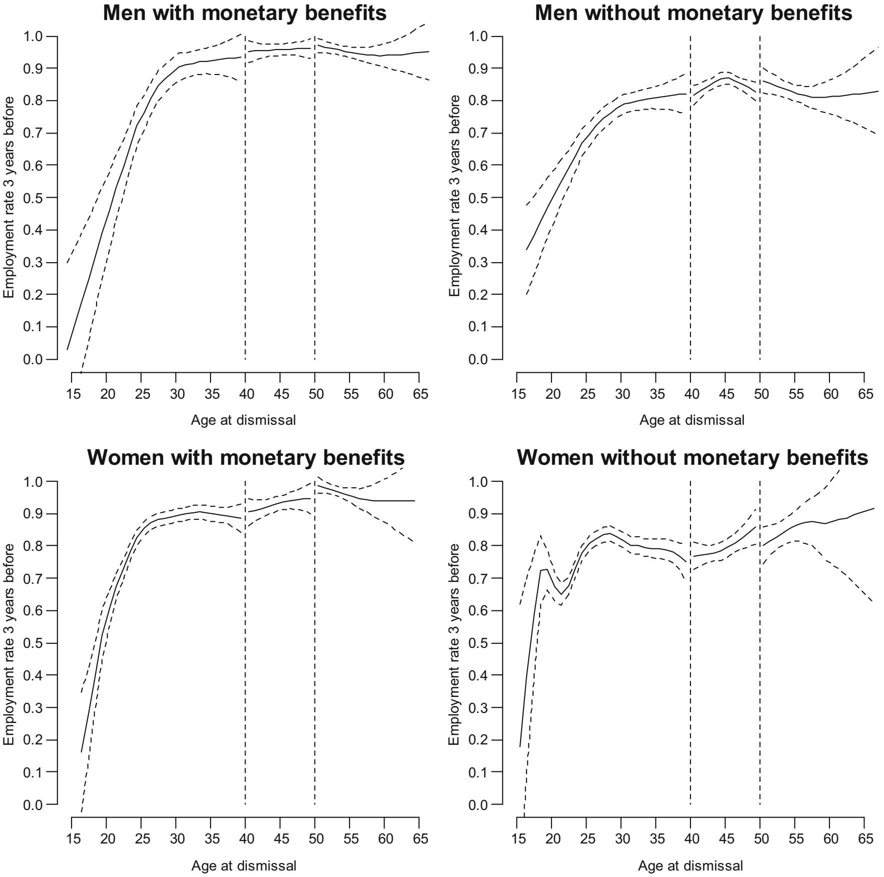


Fig. 3 Employment rate 3 years *before* enrollment in the *Liste di Mobilità* (95% confidence interval represented by the dashed lines; the vertical dashed lines represent threshold ages). Source: Paggiaro et al. (2009)

same size as that observed for wages 3 years before enrollment (not reported here; see Figure 3 in Paggiaro et al. 2009). Overall, the evidence is that the additional year of eligibility has no impact on wages 3 years after enrollment in the LM.

Summing up, the main substantive evidence on the LM program is that at 40 years, there is *no effect of the 2-year vs. 1-year eligibility regime* on reemployment rates 3 years after enrollment. There is also no effect for workers eligible only for the active component or for those eligible for the additional passive component. That is, the active component exactly counteracts the passive one. Moreover, the additional year of eligibility has no effect on wages once a new job is secured.

There is some evidence (not reported here) that women with monetary benefits postponed their reentry to work: after 2 years, the effect of the additional year of eligibility on the employment rate is −15%, which disappears 1 year later.

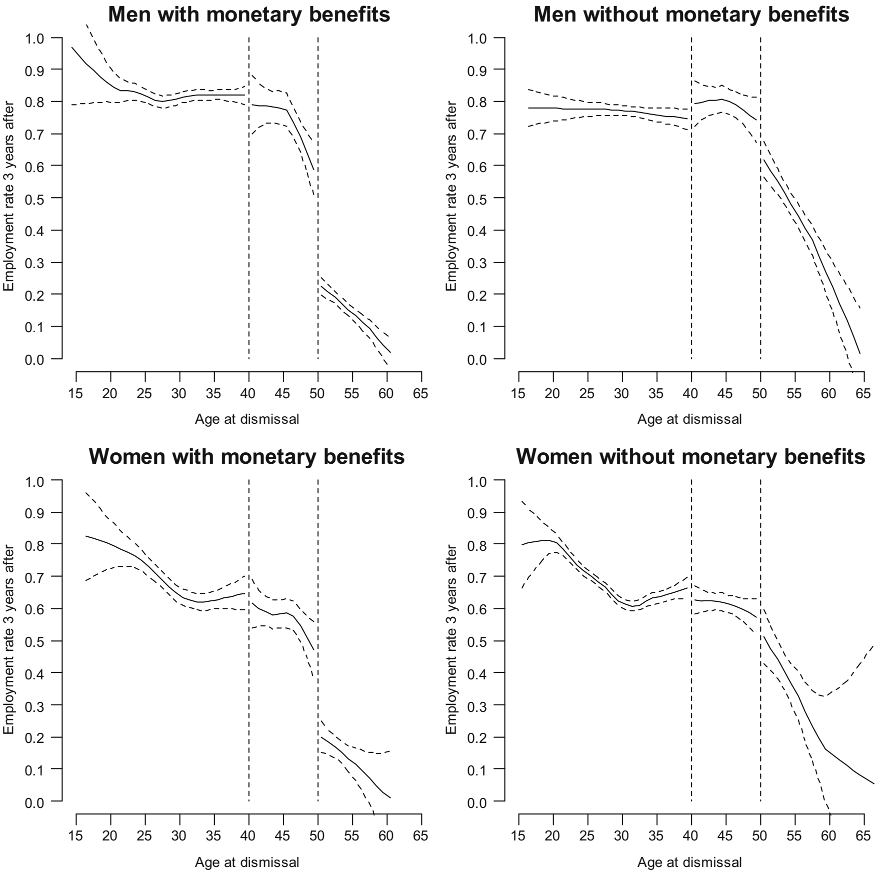


Fig. 4 Employment rates 3 years *after* enrollment in the *Liste di Mobilità* (95% confidence interval represented by the dashed lines; the vertical dashed lines represent threshold ages). Source: Paggiaro et al. (2009)

There is a strong *negative effect at 50* for workers with monetary benefits: the effect is as large as −35% and −30% for men and women, respectively. This is the effect of providing them with the option of using LM as a bridge to retirement.

4.3 Disentangling the Roles of the Active and Passive Components

Studies on the effects of the LM reviewed in the preceding sections have been motivated by the fact that, from a theoretical point of view, the sign of the effect of the 2-year regime of eligibility *vs.* the 1-year regime is a priori uncertain. Indeed, the same uncertainty is faced with respect to the effect of the "passive + transfer" component, since it is apparent that this component provides two contrasting incentives: on the one hand, the monetary benefit to the worker increases his or her reservation wage, lengthening the unemployment spell; on the other hand, the benefit transfer to the hiring firm makes the worker more likely to receive job offers.

Mazzarella et al. (2014) exploit the discontinuity along the firm size dimension to identify the causal effect of the monetary benefit provided to workers dismissed by firms above the 15-employee cutoff. Because of errors in measuring the exact status with respect to eligibility for the monetary benefit, even if the probability of receiving the monetary benefit as a function of firm size is discontinuous at 15 employees, the size of the discontinuity is smaller than 1 (see the case of workers up to age 39 in Fig. 5; there is similar evidence for older workers). This problem is dealt with by resorting to the so-called *fuzzy* RDD.

Figure 6 summarizes the results for workers aged 40 to 49. The estimated causal effect of the "passive + transfer" component on the reemployment probability is

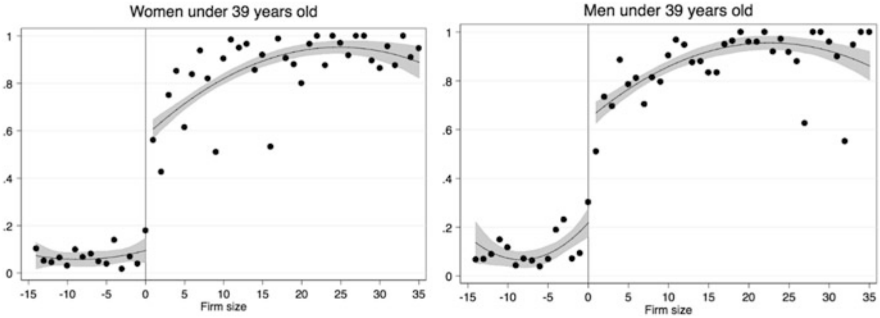


Fig. 5 Probability of receipt of the "passive + transfer" component as a function of firm size by age group and gender (point estimates and 95% confidence interval). Source: Mazzarella et al. (2014)

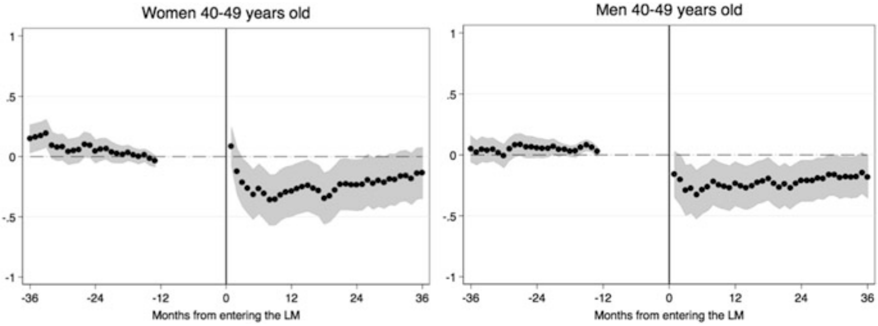


Fig. 6 Causal effect of the "passive + transfer" component on employment rate at the 15-employee threshold, from 36 months before to 36 months after enrollment in the *Liste di Mobilità* (LM), by age group and gender (point estimates and 95% confidence intervals). Source: Mazzarella et al. (2014)

nearly always *negative* over months 1 to 36 for each age–gender group, suggesting that the effect of the passive component prevails over the effect of the transfer component. The only exception is the group comprising men younger than 40. The order of magnitude of the estimated effect is around -5% for all groups except young men and is fairly stable over the 36 months.

5 Conclusions

This chapter concludes by emphasizing a collateral aspect of the story of LM and its evaluation exercises. The LM has been in operation since the early 1990s. First results on its effects started becoming available only at the end of the 1990s (Brunello and Miniaci 1997), when first-generation studies provided some estimates, albeit confined to transitions to permanent employment. Preliminary results on the impact of the LM on overall post-enrollment labor market history became available in 2005 (Paggiaro et al. 2005). Evidence of the dramatic impact of the "long mobility" clause on the labor market participation of over 50 workers became available only 4 years later, along with evidence for the lack of any impact of the duration of eligibility on earnings (Paggiaro et al. 2009). This was some *15* years after the program started its operations.

One might wonder what caused this large delay in the availability of a credible estimate of the impact of the LM. The answer is clear: it is *not* due to a lack of data. The administrative data required for the evaluation have been there since the very beginning. This means that there were conditions for obtaining a complete picture on the (mal)functioning of the program at least by the mid-1990s. The delay is first of all due to a lack of demand for evaluation on the part of policy-makers. In addition, and perhaps on a related note, the delay has been caused by the lack of an established protocol to access data for scientific purposes. Researchers took the lead in providing evaluations of LM even in the absence of explicit demand, making use of the information they were able to obtain.

There is still a missing piece of evidence on the effect of LM. It is known that at the 40-year threshold, the additional year of eligibility did not cause any improvement in the probability of securing a job nor did it boost wages once a job had been obtained. What has yet to be established is the impact of that additional year of eligibility on the costs of the program: how much did it cost to realize that zero impact on employment and wages? Again, the data to answer this question do exist. However, they were simply not accessible to researchers at the time the papers reviewed here were developed.

Acknowledgments The authors have been working on the program *Liste di Mobilità* for several years and for as long as they have been able to access rich data. During this journey they have benefited from collaboration with Adriano Paggiaro, Gianluca Mazzarella, and Nadir Zanini. Additionally, they received helpful suggestions on administrative data issues from Bruno Anastasia, Maurizio Gambuzza, and Maurizio Rasera and enjoyed stimulating discussions with Paolo Sestito. The authors' warmest thanks go to all of them.

A.1 Appendix: A 30-second Overview of the Econometrics of the LM Studies

The institutional rules of the program induce discontinuities in the level and/or type of benefits received by eligible individuals, either along the age dimension or along the firm size dimension. One way or another, all the studies reviewed in this chapter exploit those discontinuities to identify causal effects, following the logic of the regression discontinuity design (RDD) (Imbens and Lemieux 2008, Lee and Lemieux 2010). Here the RDD identification strategy is briefly reiterated, with reference to the 40-year threshold.

The duration of the eligibility period is assigned on the basis of a worker's age at dismissal only, the *running variable* in RDD terminology. Comparing workers assigned to the 2-year regime with those assigned to the alternative 1-year regime, the econometric problem is how to disentangle the causal effect of the second year of eligibility from a pure age effect. Let treatment *I*, denoting eligibility for the second year, be equal to 1 for individuals aged 40–49 and 0 for those aged up to 39.

The outcomes are the economic performances in the post-enrollment months of workers enrolled in the LM, e.g., employment status, duration of unemployment, earnings, and so on. Let Y_1 and Y_0 be the potential outcomes a specific worker would experience after being exposed to and being denied the treatment, respectively. Let Y be the outcome observed for a specific individual. It is linked to the potential outcomes and to the treatment status by the following identity:

$$Y = Y_0 + I \times (Y_1 - Y_0)\,, \tag{A1}$$

i.e., for each specific individual, either Y_0 or Y_1 is observed, depending on eligibility status.

By contrasting the average outcome experienced by the treatment group with the average outcome experienced by the comparison group, the following identity is obtained:

$$\begin{aligned} E\left[Y\,|I = 1\right] - E\left[Y\,|I = 0\right] = E\left[Y_1 - Y_0\,|I = 1\right] \\ + \left\{E\left[Y_0\,|I = 1\right] - E\left[Y_0\,|I = 0\right]\right\} \end{aligned} \tag{A.2}$$

where $E[Y_1 - Y_0|I = 1]$ is the average treatment effect on the treatment group (ATT) and the difference in brackets is the so-called selection bias, i.e., the difference observed between the two groups even in the absence of any difference in the duration of eligibility. It is induced by the differential composition of the two groups with respect to age.

By conditioning on age = 40, the only age around which there are both individuals assigned to the 2-year regime and individuals assigned to the 1-year regime, the selection bias becomes:

$$E\left[Y_0\,|I = 1\,,\,\text{age} = 40\right] - E\left[Y_0\,|I = 0\,,\,\text{age} = 40\right]. \tag{A3}$$

The classic *sharp* RDD identifying restriction states that this selection bias is zero by claiming that the conditional mean $E[Y_0 \,|\text{age}]$ is a continuous function of age in a neighborhood of age = 40. The rationale for this restriction is that, since age is by design the *only* individual characteristic relevant for the assignment of the duration of eligibility, it is the *only* individual characteristic possibly raising a selection bias problem. Hence, by comparing individuals who are very similar with respect to age, the selection bias shrinks to zero, implying that the quantity:

$$E\left[Y\,|I=1\text{ , age}\approx 40\right]-E\left[Y\,|I=0\text{ , age}\approx 40\right] \tag{A4}$$

identifies a meaningful causal parameter.

The drawback of this design is that if the program impact is heterogeneous across subjects—as is likely in most cases—then the quantity (A4) identifies the ATT only near the threshold, i.e., the RDD is useless to identify the causal effect away from the threshold value. An additional problem is that by the very nature of the causal parameter identified by the RDD—it is a causal parameter at a specific value of the running variable—only units within a certain distance of the threshold value contribute to the estimation of the causal effect. Hence, it is not the overall sample size that matters for the precision of the estimate but the number of units in a suitably defined neighborhood of the threshold value.

To estimate the ATT at the threshold value requires estimating the conditional expectations of the two potential outcomes to the left and to the right of the threshold, respectively. There are several methods available; one of the most popular is the local linear regression as proposed by Hahn et al. (2001).

A too small sample size is precisely the problem that Paggiaro and Trivellato (2002) deal with. Instead of restricting the analysis to individuals close to the threshold value (in their case, workers enrolled in the LM in an age range around 40), they use the whole sample to model the relationship between the outcome and age using a semiparametric model. Then they use the estimated model to evaluate the average outcome just above the threshold and the average outcome just below it.

The *fuzzy* RDD used by Mazzarella et al. (2014) arises when it is not the *exposure* to the treatment that jumps suddenly from 0 to 1 as the running variable crosses the threshold. Rather, it is the *probability of being exposed* that jumps suddenly but with the size of the jump smaller than 1. Mazzarella et al. (2014) observe that this happens essentially because of a measurement error problem on the running variable (see the discussion in Sect. 4.3). In this instance, the difference in average outcome between those just above the threshold (i.e., eligible for the passive component of the LM) and those just below it (i.e., not eligible for the passive component of the LM) *understates* the average causal effect because there are units below the threshold receiving the passive component as well as units above the threshold not receiving it. To correct for this bias, it is sufficient to rescale the difference in outcome across the threshold by dividing it by the jump at the threshold in the probability of exposure to the treatment. Mazzarella et al. (2014) consider it to be the jump documented in Fig. 6.

Hahn et al. (2001) show that the resulting estimate is an instrumental variable (IV) locally at the threshold value. The eligibility status—as defined by the running variable being above/below the threshold value—acts as an IV for the actual exposure to the treatment.

References

Brunello G, Miniaci R (1997) Benefit transfers in Italy: an empirical study of mobility lists in the Milan area. Oxf Bull Econ Stat 59:329–347

Hahn J, Todd P, van der Klaauw W (2001) Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69:201–209

Ham J, Lalonde R (1996) The effect of sample selection and initial conditions in duration models: evidence from experimental data on training. Econometrica 64:175–205

Imbens G, Lemieux T (2008) Regression discontinuity designs: a guide to practice. J Econ 142(2):615–635

Lee D, Lemieux T (2010) Regression discontinuity designs in economics. J Econ Lit 48(2):281–355

Lee DS (2008) Randomized experiments from non-random selection in U.S. House elections. J Econ 142(2):675–697

Mazzarella G, Rettore E, Trivellato U et al (2014) The effect of a mixed passive and active labour market policy: Evidence from an Italian programme for dismissed workers. Riv Ital Valutazione 58:80–101

McCrary J (2008) Manipulation of the running variable in the regression discontinuity design: a density test. J Econ 142(2):698–714

Mortensen D, Pissarides C (1999) New developments in models of searching the labor market. In: Ashenfelter OC, Card D (eds) Handbook of labor economics, vol 3C. Elsevier, Amsterdam, pp 2567–2627

Paggiaro A, Trivellato U (2002) Assessing the effects of the 'mobility lists' programme by flexible duration models. Labour 16:235–266

Paggiaro A, Rettore E, Trivellato U (2005) The impact of the Italian 'mobility lists' on employment chances: new evidence from linked administrative archives. Progetto MIUR 'Metodi e studi di valutazione degli effetti di politiche del lavoro, di aiuto alle imprese e di welfare', Working Paper No. 65. Department of Statistical Sciences, University of Padua, Padua

Paggiaro A, Rettore E, Trivellato U (2009) The effect of a longer eligibility to a labour market programme for dismissed workers. Labour 23(1):37–66

Enrico Rettore is Professor of Economic Statistics at the Dept. of Sociology and Social Research, University of Trento, affiliated as a senior researcher to FBK-IRVAPP (Trento), and a research fellow at IZA. His research interests cover a variety of fields of econometrics, ranging from methodological aspects to more applied work in labor economics and program evaluation. On these topics he has published in various journals, including the *American Economic Review*, the *Journal of Econometrics*, and The *Review of Economics and Statistics*. He has extensively worked on the evaluation of welfare and labor market programs in Italy. He has been the principal investigator of projects funded by the Italian Ministry of Education and Ministry of Economy and Finance, among others.

Ugo Trivellato is Emeritus Professor of Economic Statistics at the University of Padova, where he was a Professor from 1980 to 2010, FBK-IRVAPP Senior Research Fellow, IZA Fellow, and CESifo Fellow. His main research interests are program evaluation, measurement and modelling of labor supply, and unemployment. His previous research were on structural equation models with measurement errors, data revisions, and dynamic economic modelling. He is a consultant and member of advisory committees of governmental and international agencies on statistical information and microdata access, with publications in various journals, among which *European Economic Review*, *Journal of Business & Economic Statistics*, *Journal of Econometrics, Journal of the Royal Statistical Society, Labor, Politica Economica, Quality & Quantity, Rivista Internazionale di Scienze Economiche e Commerciali, Statistica,* and *Survey Methodology*.

Negative Home Equity and Job Mobility

Andrea Morescalchi, Sander van Veldhuizen, Bart Voogt, and Benedikt Vogt

1 Introduction

During the recent global financial crisis (2007–2009), the large decline in house prices led many homeowners into negative home equity (NHE). NHE occurs when the value of the house is lower than the outstanding debt on it. It has been suggested that the large increase in NHE hindered the mobility of workers and had negative consequences on the labour market (Stiglitz 2009; Krugman 2010; Katz 2014). The increase in the proportion of homeowner households with NHE was particularly high in the Netherlands. Indeed, it increased from less than 10% to more than 20% in the period 2006–2011. This increase was especially due to homeowners who fell into NHE as a result of the unexpected decline in house prices.

Several empirical studies have tested the impact of NHE on residential mobility, with mixed findings. While some studies find that NHE reduces the probability of moving (Henley 1998; Ferreira et al. 2010; Modestino and Dennett 2013; Andersson and Mayock 2014), some others find the opposite (Donovan and Schnure 2011; Schulhofer-Wohl 2011; Coulson and Grieco 2013; Bricker and Bucks 2016).

A. Morescalchi
IMT School for Advanced Studies Lucca, Lucca, Italy

European Commission Joint Research Centre (JRC), Ispra, VA, Italy
e-mail: andrea.morescalchi@imtlucca.it

S. van Veldhuizen · B. Vogt (✉)
Netherlands Bureau for Economic Policy Analysis, The Hague, Netherlands
e-mail: s.van.veldhuizen@cpb.nl; b.vogt@cpb.nl

B. Voogt
Authority for Consumer and Markets, The Hague, The Netherlands
e-mail: Bart.Voogt@acm.nl

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_12

However, whether or not homeowners with NHE, also called ‘underwater’ homeowners, are more or less mobile, only limited attention has been devoted to investigating the effect of NHE on the labour market. To our knowledge, only Mumford and Schultz (2014) have investigated the relationship between NHE and job transitions using survey data from the Panel Study of Income Dynamics (PSID). The present chapter makes use of Dutch administrative data to estimate the effect of NHE on job-to-job transitions. This is the first study investigating the impact of NHE on the labour market based on administrative data. Panel fixed-effects estimation is carried out by making use of a panel data on Dutch homeowners from the period 2006–2011.

To account for self-selection into NHE based on unobserved characteristics, the authors consider homeowners who fall into NHE because of an exogenous price decline and compare them with homeowners with positive home equity (PHE).

The impact of NHE on job mobility is expected to work via the effect on residential mobility. Although existing evidence on the effect of NHE on residential mobility is mixed, a negative effect has been found in the Dutch case (van Veldhuizen et al. 2016). Reduced propensity to relocate should render homeowners with NHE less likely to change jobs because they are more prone to discard job opportunities requiring relocation. A reduction in the probability of changing jobs may prolong inefficient job matches and deprive homeowners of interesting job opportunities that could have improved the quality of their job match (Munch et al. 2006, 2008).

The study is structured as follows: Section 2 describes the place of the study in the literature; Sect. 3 describes the dataset; Sect. 4 describes the methodological approach; Sect. 5 presents the results, including a robustness analysis; Sect. 6 provides the conclusions; and Sect. 7 reflects on policy lessons learned from this study.

2 Literature Review and Theoretical Background

The present study is related to two strands of literature. The first deals with the impact of NHE on residential mobility. The second deals with the impact of homeownership on labour market outcomes.

In the first strand of literature, there exist theories that describe either a negative or a positive impact of NHE on residential mobility. On the one hand, three reasons have been put forward to explain the well-known ‘lock-in effect’, predicting a negative effect of NHE on residential mobility. First, underwater homeowners may be less mobile because of liquidity constraints on making a down payment on a new home (Stein 1995). Second, nominal loss aversion may make underwater mortgagers less willing to sell their home after its price has fallen (Genesove and Mayer 2001; Engelhardt 2003; Cunningham and Engelhardt 2008). Third, Chan (2001) notes that the lock-in effect can be present only in the case of localised price declines. On the other hand, NHE could provide an incentive to default and hence could even increase mobility (Coulson and Grieco 2013).

Several empirical studies have tested the impact of NHE on residential mobility, with mixed findings reflecting the ambiguity of theoretical predictions. While some studies find that NHE reduces the probability of moving (Henley 1998; Ferreira et al. 2010; Modestino and Dennett 2013; Andersson and Mayock 2014), some others find the opposite (Donovan and Schnure 2011; Schulhofer-Wohl 2011; Coulson and Grieco 2013; Bricker and Bucks 2016). Ambiguity in results can be related to the fact that these studies investigate different countries, which differ in their institutional settings.

A problem with the available evidence is the potential selection of less mobile individuals into high debt levels. To the authors' knowledge, there is only one study that tries to tackle the issue of self-selection: van Veldhuizen et al. (2016) exploit exogenous switches from PHE to NHE resulting from unexpected decreases in house prices. They find that households falling into NHE are 18% less likely to move. This chapter tackles self-selection in a similar way.

In the second strand of literature, since the 1980s many scholars have maintained that homeownership should impair the functioning of the labour market.[1] This claim is based, to a large extent, on the argument that higher costs for selling and buying houses render homeowners less mobile, which has become popular under the label 'Oswald's thesis' (Oswald 1996, 1997, 1999; Blanchflower and Oswald 2013). Reduced residential mobility makes homeowners less prone to relocate for jobs, and hence they are expected to have higher reservation wages, lower search intensity and lower job-finding rates for non-local jobs but opposite outcomes for local jobs (Munch et al. 2008; Morescalchi 2016).

Empirical evidence consistently reports that homeowners are less prone to relocate for jobs (Henley 1998; Munch et al. 2006; Battu et al. 2008; van Vuuren 2009). Most microeconometric studies have found that unemployment periods of homeowners are not longer, or even shorter, than those of renters (Goss and Phillips 1997; Coulson and Grieco 2013; Flatau et al. 2003; Munch et al. 2006, 2008; Battu et al. 2008; van Vuuren 2009; Morescalchi 2016). This evidence has led to the well-known puzzle of homeowners experiencing shorter unemployment periods despite being less prone to making job-related moves (Morescalchi 2016).

Fewer microeconometric studies have investigated the impact of homeownership on transitions from employment. Evidence shows that homeownership reduces unemployment risks (van Leuvensteijn and Koning 2004; de Graaff et al. 2009; de Graaff and van Leuvensteijn 2013) as well as the likelihood of job-to-job transitions for employees. To check if lower job-to-job transitions of homeowners are explained by lower regional mobility, Battu et al. (2008) and Munch et al. (2008) break down job-to-job transitions into transitions to local jobs and transitions to jobs associated with regional relocation. They both find that homeownership reduces the likelihood of transition to non-local jobs. They also find a negative effect on transitions to local

[1] See Havet and Penot (2010) for a survey of earlier studies on the effect of homeownership on the labour market.

jobs, but this effect is smaller in both studies and not significant in the study by Battu et al. (2008).

To reconcile the argument underlying Oswald's thesis and empirical evidence, some microeconometric studies have made distinctions between outright owners and mortgagers.[2] Unemployed mortgagers should have greater incentives to search for a job to prevent foreclosure (Rouwendal and Nijkamp 2010). Consistently with this argument, unemployed mortgagers are found to have the shortest unemployment duration (Goss and Phillips 1997; Flatau et al. 2003; Brunet et al. 2007; Kantor et al. 2012) as well as the highest search intensity (Morescalchi 2016).

The two strands of literature described so far have limitations. First, investigation of the impact of NHE on residential mobility does not explicitly quantify the consequences for the labour market. Second, existing studies on the effect of housing tenure on labour market outcomes do not explicitly take into consideration the role of NHE. The present study fills these gaps by investigating the impact of NHE on the labour market. The authors are aware of only one study investigating the relationship between NHE and the labour market: Mumford and Schultz (2014) investigate the effect of NHE on the probability of becoming unemployed and on the probability of changing job using survey data from the PSID. They did not find a significant effect in either case. The present study makes use of administrative data and is based on a much larger sample, which contains more than 400,000 unique observations compared with fewer than 8000.

The impact of NHE on job mobility is expected to occur via its effect on residential mobility. Although existing evidence of the effect of NHE on residential mobility is mixed, a negative effect has been found in the Dutch case (van Veldhuizen et al. 2016). Analysis here is therefore based on the assumption that a negative relationship prevails in the Dutch case. Thus, reduced propensity to relocate can make homeowners with NHE less likely to change jobs, as they are more prone to discard job opportunities requiring relocation. The reduction in the probability of changing jobs may prolong inefficient job matches and deprive homeowners of interesting job opportunities that could improve the quality of their job match (Munch et al. 2006, 2008).

3 Dataset and Descriptive Statistics

The analysis in this chapter is based on data from 438,057 individuals followed during the period 2006–2011. This section describes the most important features of the data. The dataset is based on the full population of homeowners in the Netherlands who bought a house after 1995. This section will give an overview of the construction of the dataset and the most important descriptive statistics.

[2] Some studies have also compared social renters to private renters (McCormick 1983; Hughes and McCormick 1987; de Graaff et al. 2009; Flatau et al. 2003; Battu et al. 2008; Morescalchi 2016).

3.1 Constructing the Dataset

The dataset used for the analysis was constructed from multiple independent administrative data records, which can be obtained from Statistics Netherlands (*Centraal Bureau voor de Statistiek* (CBS)). CBS provides access to administrative datasets with information collected from various sources, such as registry records from municipalities or information on individual income based on tax files. The data are strictly confidential and may be used only for research purposes. The results may be published only after rigorous verification, which guarantees individual anonymity. The different datasets can be merged using an encrypted individual identifier, based on the individual social security number. This makes it possible to combine several sources of information on the same individual over time.[3]

In total, 17 independent administrative datasets were merged. These sets contain information on the individual's current job, address, house value and household balance-sheet information, such as income, financial assets and the value of the mortgage on the house. Information was also obtained on household composition, including number of individuals living in the household and changes in household composition, such as through marriage, divorce or registered partnership. Table 1 gives an overview of the most important datasets used in the analysis. The datasets are presented in the order used during the merging and cleaning process.

For the statistical analysis, a panel of male heads of households during 2006–2011 was used. To avoid attrition bias, all individuals who were continuously employed in all years were considered. An individual is defined as being employed in a certain year if they work for at least 10 months of the year. Robustness checks with four different employment periods are also reported.

3.2 Dependent Variable

Unique job identifiers from CBS were used at the individual level to identify job-to-job mobility.[4] An indicator variable was constructed that takes the value 1 if a job identifier changes in a given year with respect to the previous year. The job identifier was corrected for the following confounding factors: merging of companies, change of job within a company and renewal of (temporary) contracts at the same company.

[3] For a very detailed description of the construction of the dataset, see van Veldhuizen et al. (2016). This study uses their dataset and merges two additional datasets that contain labour market information for each individual, namely, the BAANKENMERKENBUS and BAANSOMMENTAB datasets.

[4] We used the variable BAANID, which is retrieved from the BAANKENMERKENBUS dataset.

Table 1 Administrative data sources and construction of the dataset

Data file	Information	Merging variable
GBAADRESOBJECTBUS	Addresses of all individuals living in the Netherlands since 1995	Person identifier and building identifier
EIGENDOMWOZ(BAG)TAB/OBJECTWONINGTAB	Type of building (owner-occupied/rental) and value of each building	Building identifier
GBAHUISHOUDENSBUS/GBAHUISHOUDENSTAB	Information on the household composition, household type, position of each person in the household and start and end date of a household with a certain composition	Person identifier
GBAPERSOONTAB	Information on date of birth, gender and place of birth	Person identifier
IVB, IHI	Information on household wealth, income, debt, mortgage and financial assets at the household level	Person identifier
IPI	Information on gross personal income and income source	Person identifier
VSLGWBTAB	Information on location (municipality and postcode area) of each building	Building identifier
BAANKENMERKENBUS/BAANSOMMENTAB	Information about periods of employment, type of job and sector	Person identifier and job identifier

3.3 Independent Variables

The dataset contains rich information on the balance sheet of the households. Information was obtained on the outstanding mortgage and the current value of the house on 1 January in each year. The outstanding value of the mortgage was extracted from administrative tax records. The value of the property was obtained using the official valuation of property (*waardering onroerende zaken* (WOZ) value), which is estimated by the municipality. Each household in the Netherlands receives a letter every year with information on the current value of the property. This is an approximation for the actual market value of the house. The ratio of the transaction price and the WOZ value in the sample period was 99% (CBS 2014).[5]

The mortgage loan and the value of the house were used to calculate loan-to-value ratio (LTV). Individuals were defined as having NHE if the LTV exceeded 100. To capture the effect of NHE, outcomes were compared between underwater mortgagers and mortgagers with LTVs of less than 100. However, the division between the two states may depend on unobserved heterogeneity. Therefore, two types of underwater mortgagers were defined. More specifically, the authors distinguished between voluntarily and involuntarily underwater households. Individuals can opt into a high LTV by getting a very large mortgage. An increase in debt may reflect unwillingness to move in the foreseeable future. In this case, the straight comparison between underwater mortgagers and those with PHE may simply capture a different propensity for mobility rather than an impact of NHE.

Individuals were defined as having involuntary NHE if they experienced a decline in their house value that was sufficiently large to increase the LTV to above 100. In cases where an individual's mortgage went underwater because of a combination of an increase in the mortgage and a reduction in the house price, they were considered to have voluntary NHE only if the ratio between the current mortgage and the value of the year before was also above 100. For individuals who were underwater in the first year of the sample, this distinction cannot be used. However, if they bought a house in the first year and the LTV exceeded the cutoff, they were allocated to the category of those with voluntary NHE. Once an individual was categorised as underwater, they were defined as being underwater in the same category as long as the LTV remained above 100. If the LTV fluctuated above or below 100, the underwater status was updated in accordance with these rules.

[5] This information was obtained from CBS (2014), p. 8, Table 2.3.2.2. This is the average ratio of the yearly average transaction price and the yearly average WOZ value in the sample period from 2006 to 2011. To calculate this ratio, the reference date in $t + 1$ was used as the value is determined on 1 January each year. Hence the corresponding ratio for the year 2006 is the reference date in 2007.

To capture the effect of NHE, the relevant comparison was made between mortgagers who were involuntarily underwater and those with PHE. To distinguish between voluntarily and involuntarily underwater mortgagers, two binary indicators were included in the regressions.

The data also contain a large set of control variables. In particular, use was made of housing tenure in years, household size, disposable real household income, real financial assets, household composition and changes in household composition, year and 40 regional indicators (local labour market level, *COördinatie Commissie Regionaal OnderzoeksProgramma* (COROP) regions) as control variables.

3.4 Descriptive Statistics

Table 2 shows descriptive statistics for each year. The table shows means of the variables and the corresponding standard errors in parentheses below. There are two important messages from the descriptive statistics. First, there is a general decline in job mobility in the sample period. In 2006 about 8.18% of the individuals in the sample changed job, whereas only 4.39% in 2011. Second, a striking "exogenous" increase in home equity status was observed. In 2006, only 0.24% of the households had involuntary NHE. This number increased to 9.7% in 2011. In the same period, the number of households with voluntary NHE increased only slightly (8.6% in 2006 compared with 10.93% in 2011).

4 Methodology

In estimating the impact of NHE, the authors took into consideration the fact that the NHE status can be related to unobserved characteristics that in turn may have an impact on the outcome variable. This issue was tackled using three measures.

First, a panel fixed-effects method was employed to remove the potential endogeneity bias arising from time-constant unobserved heterogeneity. Second, a set of control variables was introduced to account for time-varying heterogeneity. Third, the group of homeowners with NHE was divided into the following two groups: mortgagers who fell into NHE (1) because of a house price decline and (2) because of a voluntary increase in the mortgage loan. To capture the impact of NHE, group (1) was compared with homeowners with PHE. In this way the assignment to one of the two categories was determined exogenously.

Table 2 Descriptive statistics

	Year					
Variable	2006	2007	2008	2009	2010	2011
Job mobility	0.0818	0.0710	0.0501	0.0424	0.0498	0.0439
	(0.0004)	(0.0004)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
No NHE	0.9112	0.9115	0.8990	0.8707	0.8400	0.7937
	(0.0004)	(0.0004)	(0.0005)	(0.0005)	(0.0006)	(0.0006)
Involuntary NHE	0.0024	0.0062	0.0139	0.0358	0.0587	0.0970
	(0.0001)	(0.0001)	(0.0002)	(0.0003)	(0.0004)	(0.0004)
Voluntary NHE	0.0863	0.0823	0.0871	0.0935	0.1013	0.1093
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0005)	(0.0005)
Disposable HH income	39,614.3555	41,743.5703	42,503.1172	43,716.8555	43,584.4336	43,579.7617
	(25.4599)	(28.6924)	(28.6641)	(28.6754)	(28.8234)	(29.2537)
Financial assets	49,149.0313	52,177.6914	49,252.9688	54,359.4648	54,927.9375	58,109.3789
	(246.4086)	(257.5025)	(393.0074)	(431.23)	(283.5162)	(452.015)
Mortgage	174,023.4219	175,338.6563	174,457.4375	174,089.0000	172,893.3438	168,693.0313
	(134.4507)	(136.0783)	(136.9659)	(137.5357)	(137.703)	(137.3129)
House price	273,946.1563	279,334.0313	274,831.1875	265,920.9688	256,977.1250	242,236.5469
	(173.1247)	(177.928)	(175.7459)	(169.3979)	(162.8502)	(154.1337)
Loan age	5.5783	6.4085	7.2482	8.1548	9.0618	9.9788
	(0.0051)	(0.0052)	(0.0054)	(0.0056)	(0.0057)	(0.0059)
Married	0.7054	0.7163	0.7251	0.7313	0.7357	0.7380
	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)
Partner	0.1820	0.1715	0.1626	0.1552	0.1487	0.1438
	(0.0006)	(0.0006)	(0.0006)	(0.0005)	(0.0005)	(0.0005)

(continued)

Table 2 (continued)

Variable	Year					
	2006	2007	2008	2009	2010	2011
Number of children	1.2922	1.3249	1.3487	1.3618	1.3661	1.3605
	(0.0017)	(0.0017)	(0.0017)	(0.0017)	(0.0017)	(0.0017)
HH size	3.1955	3.2293	3.2538	3.2657	3.2676	3.2590
	(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0019)
Age	40.8809	41.8809	42.8809	43.8809	44.8809	45.8809
	(0.0117)	(0.0117)	(0.0117)	(0.0117)	(0.0117)	(0.0117)
Divorce	0.0050	0.0057	0.0059	0.0061	0.0065	0.0066
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
N	438,057					

Notes. The table shows means and standard errors (in parentheses) for each year of the panel of all variables that were used in our analysis. ‘Job mobility’ indicates the proportion of household heads that changed jobs in a subsequent year. ‘PHE’ shows the proportion of households that have an LTV lower than 100%. The variable ‘Involuntary NHE’ shows the proportion of households that fell into NHE because of a fall in the house price. ‘Voluntary NHE’ shows the proportion of households that have NHE because of an increase in the mortgage. All monetary variables such as disposable household (HH) income, financial assets, mortgage and house price are converted to 2011 euros. ‘Loan age’ indicates the years an individual has been living in the current house. ‘Married’ and ‘Partner’ are dummy variables that take the value 1 if an individual is married or in a registered partnership

HH household, *NHE* negative home equity

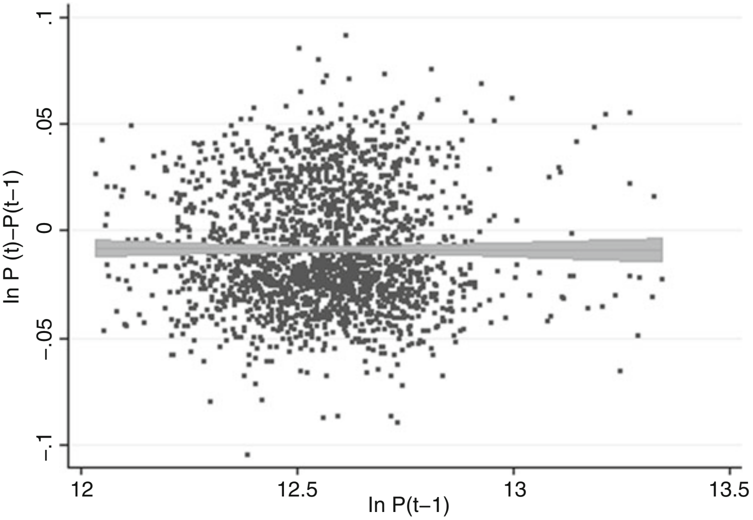


Fig. 1 No relationship between change and level of average house prices in municipalities. Notes. The grey line shows a linear regression with 95% confidence intervals

One crucial assumption of the identification was that the change in house prices was random and unexpected. Figure 1 shows the relationship between the growth rate and the level of the average house price per municipality in a given year for the period 2006–2011. The picture that emerges from this figure is that there is no relationship between the two variables. Coefficients of year-by-year regressions are also very close to zero and generally non-significant. This evidence suggests that house prices follow a random walk process, providing support to the assumption made here.

Using the sample of all employed homeowners in year t, job-to-job transitions can be modelled by the outcome variable y_{t+1}, indicating a change of job in the following year. Hence, the following linear probability model is estimated as:

$$y_{it+1} = \alpha_i + \beta 1\,(\text{InvoluntaryNHE}_{it}) + \delta 1\,(\text{VoluntaryNHE}_{it}) + \gamma X_{it} + \epsilon_{it} \quad (1)$$

In Eq. (1), β is the main coefficient of interest. α_i is an individual fixed effect, which captures all time-invariant unobserved individual heterogeneity. The parameter vector γ captures the effects of other observable time-varying characteristics that are summarised in the matrix X_{it} (X_{it} contains also a constant). It contains housing tenure, disposable household income and household financial assets. Indicator variables were also included for partnership status, marriage status and divorce status, as well as indicator variables for household size. In addition, year dummies, region dummies and region dummies interacted with year dummies were included to control for potential local labour market shocks, which can vary over time. ϵ_{it} is an error term with the usual assumptions.

5 Results

5.1 Negative Home Equity and Job-to-Job Mobility

The effect of NHE on job-to-job mobility was analysed. Estimates of parameters in Eq. (1) are reported in Table 3. Coefficients in Table 3 have to be interpreted in terms of percentage point changes of the probability of changing jobs in the subsequent year.

Column 1 reports the results of the baseline specification. This specification contains year dummies and dummies for housing tenure. The main variable of interest is the home equity status of the household. This is a categorical variable reflecting three states defined as follows. The first is mortgagers who are underwater because of a decline in the house price. They are considered to be involuntarily underwater. The effect of NHE is captured by the coefficient of this indicator. The second category corresponds to mortgagers who are underwater because they deliberately chose a high LTV either at the very outset of purchasing their home or as a result of an increase in the mortgage. They are considered to be voluntarily underwater. The third and baseline category describes the situation in which a household has PHE.

Column 1 shows that NHE reduces the probability of changing jobs in the following year by 0.339 percentage points. The results remain nearly unchanged when we add further control variables. In column 2 the study controls for disposable household income, financial assets and household size.[6] The estimates of household income show a negative association between higher household incomes and the propensity to change job in the consecutive year. Higher financial assets show a modest but mostly significant negative association with job mobility.

The results do not change if further controls are added for the change in household composition in column 3. Namely, three indicator variables are added that capture the household composition and changes to the household composition. A dummy variable is included that takes the value 1 if a household head is living in a registered partnership or marriage. The other two indicator variables take the value 1 if a household head is divorced in year *t* or is going to divorce in the following year.

The results also remain unchanged if the study controls for local labour market conditions. Since job mobility patterns can differ between labour markets, control variables are added for local labour market conditions. The Netherland is divided into 40 local labour markets, named COROP regions. Region dummies as well as their interactions with year dummies are added in columns 4–6 to control for differences in labour markets and yearly local shocks, respectively. The same sequence of regressions is in columns 1–3, so, for example, column 4 contains the same control variables as column 1. All regressions reveal point estimates that are very similar to the initial specification in column 1. Column 6 shows that plunging into NHE is associated with a 0.295 percentage point decrease in the probability of changing jobs. Since the average probability of changing jobs across all years is 5.65%, this boils down to a relative effect of about 5.2% ($\frac{0.00295}{0.0565} \times 100$).

[6] All financial variables are converted to 2011 euros.

Table 3 Determinants of job-to-job mobility

	1	2	3	4	5	6
No NHE	Reference category					
Involuntary NHE	−0.00339**	−0.00294**	−0.00290*	−0.00345**	−0.00299**	−0.00295**
	(0.00112)	(0.00112)	(0.00112)	(0.00113)	(0.00113)	(0.00113)
Voluntary NHE	−0.00235	−0.00221	−0.00219	−0.00246	−0.00231	−0.00229
	(0.00138)	(0.00138)	(0.00138)	(0.00138)	(0.00138)	(0.00138)
HH income > 20,000€ and HH income ≤ 30,000€		−0.00929**	−0.00948**		−0.00936**	−0.00955**
		(0.00161)	(0.00161)		(0.00161)	(0.00161)
HH income > 30,000€ and HH income ≤ 40,000€		−0.0143**	−0.0146**		−0.0144**	−0.0147**
		(0.00173)	(0.00173)		(0.00173)	(0.00173)
HH income > 40,000€ and HH income ≤ 50,000€		−0.0186**	−0.0189**		−0.0188**	−0.0190**
		(0.00183)	(0.00183)		(0.00183)	(0.00183)
HH income > 50,000€		−0.0215**	−0.0218**		−0.0215**	−0.0218**
		(0.00193)	(0.00193)		(0.00193)	(0.00193)
Financial assets > 10,000€ and financial assets ≤ 20,000€		−0.000761	−0.000802		−0.000721	−0.000762
		(0.000632)	(0.000632)		(0.000632)	(0.000632)
Financial assets > 20,000€ and financial assets ≤ 30,000€		−0.00164*	−0.00169*		−0.00158*	−0.00164*
		(0.000792)	(0.000792)		(0.000792)	(0.000793)
Financial assets > 30,000€ and financial assets ≤ 40,000€		−0.00239**	−0.00248**		−0.00235*	−0.00244**
		(0.000926)	(0.000926)		(0.000926)	(0.000926)
Financial assets > 40,000€		−0.00186*	−0.00196*		−0.00180	−0.00191*
		(0.000936)	(0.000937)		(0.000936)	(0.000937)
Divorced			−0.00946**			−0.00940**
			(0.00193)			(0.00193)

Partner/married			0.00630**			0.00625**
			(0.00140)			(0.00140)
Will divorce			−0.00398*			−0.00398*
			(0.00181)			(0.00181)
Other controls	No	Yes	Yes	No	Yes	Yes
Region dummies and region (×) year dummies	No	No	No	Yes	Yes	Yes
Observations	2,628,342					
Number of individuals	438,057					

Notes. The table shows results from a linear panel regression with fixed effects at the individual level. The dependent variable takes the value 1 if an individual changes jobs in the subsequent year. All regressions contain year-fixed effects and controls for housing tenure. The variables household (HH) income and financial assets are converted to 2011 euros. ‘Other controls’ contain dummy variables for the household size. Robust standard errors are in parentheses

HH household, *NHE* negative home equity

* $p < 0.01$, * $p < 0.05$

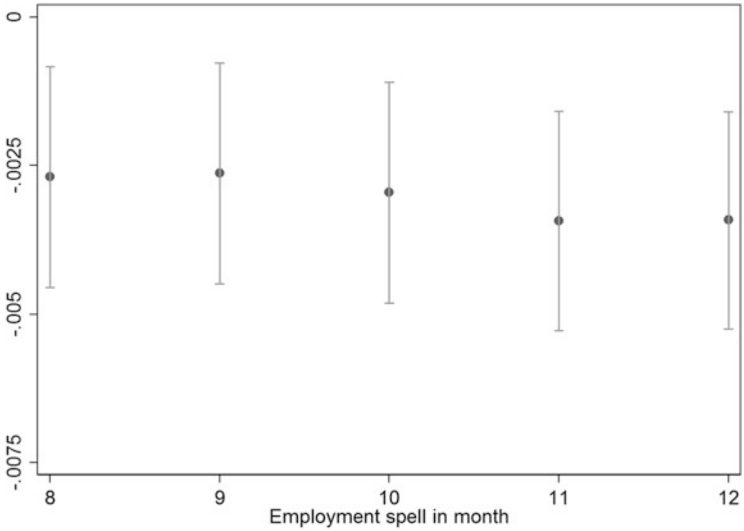


Fig. 2 Effect of involuntary negative home equity on job-to-job mobility for different periods of employment. Notes. The figure shows point estimates with 95% confidence intervals of linear panel regression, with job-to-job mobility as a dependent variable for different samples. The *x*-axis indicates the minimum employment length per year. The point estimate for an employment length of 10 months is from column 6 in Table 2. The full regression results are available upon request

Table 3 also shows no relation between voluntary NHE and job-to-job mobility. None of the coefficients in columns 1–6 are significantly different from zero.[7]

5.2 Robustness

A crucial selection was made with regard to the sample about the length of employment. The main analysis focused on an employment period of at least 10 months in each year. One important question was hence to check if the results changed if longer or shorter employment periods were allowed.

To address this question, the same analysis was conducted with samples based on different employment periods. The authors looked at periods of 8, 9, 11 and 12 months of employment in each year and ran the same analysis as in Table 3.

The analysis reveals that the choice of the length of the employment period does not influence the main results. Figure 2 shows the point estimates with 95% confidence intervals of the key variable of interest. The *x*-axis shows the length

[7] As an alternative estimation, we also run an instrumental variable (IV) regression, pooling the category of voluntary NHE with the reference category of no NHE. The binary indicator for NHE has been instrumented with residuals of a house price equation containing the full set of regressors. The coefficient of the binary indicator is not significant in the second stage, consistent with the fact that voluntary and involuntary NHE are associated with similar effects in Table 3. Results are available from the authors on request.

of employment in each year. Regardless of the choice of employment period, a statistically significant and negative relation is found between plunging into NHE and the propensity to change jobs. The figure also shows that the point estimates for each employment period are not statistically significantly different from each other. NHE is associated with a moderate decrease in job-to-job mobility.[8]

6 Summary

This study investigated the effect of NHE status on job-to-job mobility. It used a Dutch administrative panel dataset of homeowners for the period 2006–2011. The analysis reveals a small negative effect of NHE on job-to-job mobility. Households that plunge into NHE involuntary as a result of an unforeseen fall in house prices are 5.2% less likely to change jobs than households remaining with PHE. No relationship is found between voluntary NHE and job-to-job mobility.

The effect of NHE on job-to-job mobility is relatively small, especially if compared with the effect on household residential mobility. There are three main reasons why the effect might be so small. First, the institutional setup, such as the National Mortgage Insurance Scheme, covers the mortgage payments in case of default due to involuntary unemployment. This mitigates the potential risks of changing jobs as a homeowner.

Second, in the Netherlands, high levels of mortgage debt are not associated with weak borrower characteristics or higher default probabilities (Mocking and Overvest 2015). This is one of the key differences from the United States and some other countries, where high levels of mortgage debt are often associated with weaker borrower characteristics, such as low income, low socio-economic background or low educational levels. When the housing market collapsed, many of these 'subprime' mortgages defaulted because many of the mortgagers became unemployed as a result of their weaker positions in the labour market.

Third, homeowners with NHE might be prone to longer commuting distances. In this case they might take up new opportunities in distant areas almost as often as homeowners with positive equity, but avoiding relocation. The metropolitan region 'Randstad' covers the four biggest cities in the country: Amsterdam, Rotterdam, The Hague and Utrecht. In this region the infrastructure is such that individuals do not have to move when they change jobs. The impact of NHE on commuting is a topic for future research.

[8] The full set of regression results of this robustness check is available upon request.

7 Policy Lessons

The reduced labour mobility of homeowners plunging into NHE can have adverse effects on the economy through reduced labour market matching efficiency, which can lead to lower productivity levels. Two types of policies can be identified to reduce these adverse effects: policies targeted at homeowners facing negative equity and policies targeted at the flow of future homeowners with low equity levels.

Policies regarding homeowners facing negative equity could be aimed at reducing barriers to refinancing of residual debt. In general, mortgage lenders typically do not (re)finance high LTVs. In the United States, most mortgage lenders will not finance over 80% loan-to-home value, whereas in the Netherlands the cutoff is 101%. Policies to refinance underwater mortgages could be government-run programmes that offer high LTV loans, like the Home Affordable Refinance Program (HARP) in the United States, which offers loans with LTVs up to 125% to cover the home's reduced value. In addition to such government-run programmes, another policy could be to make the interest paid on the refinanced debt tax deductible. This is done in the Netherlands, for example, to incentivise households to accelerate repayments of residual debt. Finally, governments could reduce transaction costs by lowering stamp duty on residential moves that accompany job changes.

Policies aimed at the future flow of homeowners with low equity levels are typically targeted at reducing the incentives to enter into high mortgage debts. This can be achieved by lowering mortgage interest-rate deductions, which are typically found in some advanced economies (such as the United States and the Netherlands). A home mortgage interest deduction allows taxpayers who own their homes to reduce their taxable income by the amount of interest paid on the loan that is secured by their principal residence. An additional regulatory policy could be (the introduction of) stricter mortgage standards, such as limits on loan-to-value or loan-to-income ratios. These macroprudential tools are frequently considered to dampen credit growth and avoid boom and bust cycles on housing markets. However, the institutional setting should be taken into consideration when planning restrictions in the availability of credit. In the Netherlands, for example, households face relatively large mandatory pension savings that restrict any additional savings for most households, which could lead to reduced access to the housing market (van Veldhuizen et al. 2015).

Acknowledgements The authors would like to thank Michiel Bijlsma, Jonneke Bolhaar, Sander Gerritsen and Remco Mocking for helpful comments. The codes used to construct the dataset and to conduct the analysis are available upon request. The data used for analysis are strictly confidential and available only for personally registered users of Statistics Netherlands (CBS-remote access). The authors will provide all necessary assistance to gain access to the data. All remaining errors are their own.

References

Andersson F, Mayock M (2014) How does home equity affect mobility? J Urban Econ 84:23–39

Battu H, Ma A, Phimister E (2008) Housing tenure, job mobility and unemployment in the UK. Econ J 118(527):311–328

Blanchflower DG, Oswald AJ (2013) Does high home-ownership impair the labor market? NBER Working Paper No 19079. National Bureau of Economic Research, Cambridge

Bricker J, Bucks B (2016) Negative home equity, economic insecurity, and household mobility over the Great Recession. J Urban Econ 91:1–12

Brunet C, Clark AE, Lesueur J-Y (2007) Statut résidentiel et durée de chômage en France et au Royaume-Uni. Rev Fr Econ 22(2):165–190

CBS (Statistics Netherlands) (2014) Prijsindex Bestaande Koopwoningen: methodebeschrijving. CBS, The Hague

Chan S (2001) Spatial lock-in: do falling house prices constrain residential mobility? J Urban Econ 49(3):567–586

Coulson NE, Grieco PLE (2013) Mobility and mortgages: evidence from the PSID. Reg Sci Urban Econ 43(1):1–7

Cunningham CR, Engelhardt GV (2008) Housing capital-gains taxation and homeowner mobility: evidence from the Taxpayer Relief Act of 1997. J Urban Econ 63(3):803–815

de Graaff T, Van Leuvensteijn M (2013) A European cross-country comparison of the impact of homeownership and transaction costs on job tenure. Reg Stud 47(9):1443–1461

de Graaff T, Van Leuvensteijn M, van Ewijk C (2009) Homeownership, social renting and labor mobility across Europe. In: van Ewijk C, Van Leuvensteijn M (eds) Homeownership and the labour market in Europe. Oxford University Press, Oxford, pp 53–81

Donovan C, Schnure C (2011) Locked in the house: do underwater mortgages deduce labor market mobility? https://doi.org/10.2139/ssrn.1856073

Engelhardt GV (2003) Nominal loss aversion, housing equity constraints, and household mobility: evidence from the United States. J Urban Econ 53(1):171–195

Ferreira F, Gyourko J, Tracy J (2010) Housing busts and household mobility. J Urban Econ 68(1):34–45

Flatau P, Forbes M, Hendershott PH (2003) Homeownership and unemployment: the roles of leverage and public housing. NBER Working Paper No 10021. National Bureau of Economic Research, Cambridge

Genesove D, Mayer C (2001) Loss aversion and seller behavior: evidence from the housing market. Q J Econ 116(4):1233–1260

Goss EP, Phillips JM (1997) The impact of home ownership on the duration of unemployment. Rev Reg Stud 27(1):9

Havet N, Penot A (2010) Does homeownership harm labour market performances? A survey. doi:https://doi.org/10.2139/ssrn.1625248

Henley A (1998) Residential mobility, housing equity and the labour market. Econ J 108(447):414–427

Hughes G, McCormick B (1987) Housing markets, unemployment and labour market flexibility in the UK. Eur Econ Rev 31(3):615–641

Kantor Y, Nijkamp P, Rouwendal J (2012) Homeownership, unemployment and commuting distances. doi:https://doi.org/10.2139/ssrn.2327010

Katz L (2014) Long-term unemployment in the Great Recession. Harvard University, Cambridge

Krugman P (2010) Beveridge worries. New York Times, 29 July 2010. http://krugman.blogs.nytimes.com/2010/07/29/beveridge-worries/?_r=0. Accessed 24 Jul 2017

McCormick B (1983) Housing and unemployment in Great Britain. Oxford Econ Pap 35:283–305

Mocking R, Overvest B (2015) Estimating the impact of forced sales on house prices. CPB Discussion Paper. CPB Netherlands Bureau for Economic Policy Analysis. http://econpapers.repec.org/paper/cpbdiscus/304.htm. Accessed 24 Jul 2017

Modestino AS, Dennett J (2013) Are American homeowners locked into their houses? The impact of housing market conditions on state-to-state migration. Reg Sci Urban Econ 43(2):322–337

Morescalchi A (2016) The puzzle of job search and housing tenure: a reconciliation of theory and empirical evidence. J Reg Sci 56(2):288–312

Mumford KJ., Schultz K (2014) The effect of underwater mortgages on unemployment. Columbia University, New York. http://www.krannert.purdue.edu/faculty/kjmumfor/papers/Underwater_and_Unemployed.pdf. Accessed 24 Jul 2017

Munch JR, Rosholm M, Svarer M (2006) Are homeowners really more unemployed? Econ J 116(514):991–1013

Munch JR, Rosholm M, Svarer M (2008) Home ownership, job duration, and wages. J Urban Econ 63(1):130–145

Oswald AJ (1996) A conjecture on the explanation for high unemployment in the industrialized nations: part 1. University of Warwick, Coventry

Oswald AJ (1997) Thoughts on NAIRU. J Econ Perspect 11(4):227–228

Oswald AJ (1999) The housing market and Europe's unemployment: a non-technical paper. University of Warwick, Coventry

Rouwendal J, Nijkamp P (2010) Homeownership and labour-market behaviour: interpreting the evidence. Environ Plan A 42(2):419–433

Schulhofer-Wohl S (2011) Negative equity does not reduce homeowners' mobility. NBER Working Paper 16701. National Bureau of Economic Research, Cambridge. http://www.nber.org/papers/w16701. Accessed 24 Jul 2017

Stein JC (1995) Prices and trading volume in the housing market: a model with down-payment effects. Q J Econ 110(2):379–406

Stiglitz JE (2009) The challenge of creating jobs in the aftermath of the 'Great Recession.' Testimony before the Joint Economic Committee, Washington, DC, 10 Dec 2009

Van Leuvensteijn M, Koning P (2004) The effect of home-ownership on labor mobility in the Netherlands. J Urban Econ 55(3):580–596

van Veldhuizen S, Groot S, van Dijk M (2015) De economische effecten van een verdere verlaging van de LTV-limiet (The economic effects of a further reduction of the LTV limit). CPB Discussion Paper. CPB Netherlands Bureau for Economic Policy Analysis, The Hague

van Veldhuizen S, Vogt B, Voogt B (2016) Negative home equity and household mobility: evidence from administrative data. CPB Discussion Paper No 323. CPB Netherlands Bureau for Economic Policy Analysis, The Hague. https://ideas.repec.org/p/cpb/discus/323.html. Accessed 24 Jul 2017

van Vuuren, A (2009) The impact of homeownership on unemployment in the Netherlands. In: van Ewijk C, van Leuvensteijn M (eds) Homeownership and the labour market in Europe. Oxford University Press, Oxford, pp 113–136

Andrea Morescalchi is a researcher in Economics working for the European Commission Joint Research Centre (Ispra, Italy). He holds a degree and a PhD in Economics from the University of Pisa. After the PhD he worked as post-doc fellow at IMT Lucca and then moved to the JRC. His main research interests are the interplay between labour, housing and regional economics, impact evaluation, networks of knowledge and innovation.

Sander van Veldhuizen is heading the research programme public finance at the CPB Netherlands Bureau for Economic Policy Analysis, the Dutch public think-tank for economic policy research. He obtained a PhD in Applied Mathematics at Delft University of Technology. In the period 2014–2017, he headed the financial markets research programme at CPB. Before joining the CPB, he worked as a strategist at the Netherlands Authority for Financial Markets (AFM). Sander van Veldhuizen has extensive experience in applied economic research and policy analysis in various field, amongst others financial markets, housing and innovation.

Bart Voogt is working at Authority for Consumers and Markets (ACM) as senior economist. He is involved in the regulation of the Dutch Telecommunications, Transport and Postal sectors. In 2012 he graduated from the Erasmus University Rotterdam with a PhD in Economics. His main research interests are industrial organisation, the housing market and regulated markets.

Benedikt Vogt is an applied economist with a broad field of interest for policy-relevant questions. Since 2014, he has been working as a researcher at the department of market regulation at the CPB Netherlands Bureau for Economics Policy Analysis. He graduated from the University of Bonn in 2011, and he holds a PhD in Economics of Maastricht University since 2015. His research interests lay in empirical household finance, behavioural economics and experimental economics.

Microdata and Policy Evaluation at CPB

Rudy Douven, Laura van Geest, Sander Gerritsen, Egbert Jongen, and Arjan Lejour

1 The Activities of CPB Netherlands Bureau for Economic Policy Analysis

CPB Netherlands Bureau for Economic Policy Analysis was established in 1945. CPB is the acronym for the Dutch name *Centraal Planbureau*. CPB is an independent economic think tank, funded by the government, and aims to provide independent, impartial economic analysis that is policy relevant and academically up to standard. CPB's main challenge is to contribute to better economic policies, predominantly through applied empirical research. Theory and good intentions alone are not enough to ensure effective and efficient policy. It is important to establish whether or not these policies will really bear fruit. According to a Dutch saying, 'measurement is the key to knowledge'.

R. Douven
CPB Netherlands Bureau of Economic Policy Analysis, The Hague, The Netherlands

Erasmus University, Rotterdam, The Netherlands
e-mail: R.C.M.H.Douven@cpb.nl

Laura van Geest · Sander Gerritsen
CPB Netherlands Bureau of Economic Policy Analysis, The Hague, The Netherlands
e-mail: L.B.J.van.Geest@cpb.nl; S.B.Gerritsen@cpb.nl

Egbert Jongen
CPB Netherlands Bureau of Economic Policy Analysis, The Hague, The Netherlands

Leiden University, Leiden, The Netherlands
e-mail: E.L.W.Jongen@cpb.nl

Arjan Lejour (✉)
CPB Netherlands Bureau of Economic Policy Analysis, The Hague, The Netherlands

Tilburg University, Tilburg, The Netherlands
e-mail: A.M.Lejour@cpb.nl

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_13

Over time, research at CPB has moved from mainly macroeconomic research to a more balanced approach between macro and micro. With the increased accessibility of microdata and the rising power of computers, new research avenues have opened up. At CPB, various routes are explored to enhance evidence-based policies using microdata.

First of all, CPB conducts microeconometric analysis, over a wide range of topics including taxation, education, health care, labour market policies and wealth and income inequality. Academic papers, simulation models and policy briefs are the main outlets. CPB's value added is the focus on the Dutch context, in terms of the data used, the underlying institutions and the choice of research topics.

CPB tends to use microdata that are readily available. Occasionally, CPB collects or compiles its own microdata, undertakes surveys or conducts experiments. Statistics Netherlands, the Dutch Health Care Authority, the Tax and Customs Administration Office and the Dutch Central Bank are CPB's main data providers.

CPB sometimes advises Dutch government ministries on the design of evaluations or experiments undertaken by others: what are viable routes forward? On other occasions CPB draws up guidelines. These guidelines are often used by ministries when they tender evaluations or experiments. In this way, CPB ensures that research projects set out by ministries meet a certain minimum standard.

CPB also assesses new policy proposals, using available international microeconometric research. Governments may want an assessment of policies before their actual implementation. In the absence of reliable Dutch data, it may still be possible to make an assessment by using information from international research. CPB followed this approach, for example, when the government was contemplating various options for an increase in the minimum wage for young people.

Box 1 'Promising Policies'—A Toolbox for Policy-Makers

Aim: the series 'Promising Policies' aims to foster evidence-based policies. Insight into the effectiveness of policies helps policy-makers to make better policy choices. Academic research becomes more accessible for policy-makers in the form of a practical toolbox, drawing from existing literature or models.

Format: a typical report contains (1) an overview of the state of knowledge in a certain policy area, in terms of both outcomes and existing policies, benchmarked against international peers, and (2) a series of policy options. The options are summarized in a table. The text is geared towards policy-makers and the public at large, rather than academics.

Choice of topics: a broad range of topics, on the intersection of political priorities and expertise of the agencies, are involved. The choice of topics should be validated to avoid the appearance of political bias. The series contains reports on the labour market, innovation, education, mobility, science policy and housing.

(continued)

Choice of policies: policy options are distilled from academia, the public debate (politicians, ministries, social partners, civil society) and international peers. Reports provide an overview of options that cater to a wide political spectrum.

Choice of indicators: trade-offs are the rule in economics. The reports provide scores on various quantitative indicators and often also a qualitative indicator, enabling policy-makers to weigh the various pros and cons of a policy.

An example: the reports on labour market policies (CPB 2015, 2016a) include policy options on fiscal policy, social security, employment protection, retirement age and old age pensions, wage formation and active labour market policies. The table reports effects on budgetary impact, employment, productivity, income distribution and miscellaneous factors.

More generally, CPB tries to bridge the gap between academia and policy-makers by making academic research accessible to a broader audience. Therefore CPB uses various channels: policy reports, scientific publications and books and seminars and presentations suited for various audiences. A new series, ‘Promising Policies’,[1] aims to provide an overview of reform options in the Dutch situation for certain policy areas (see Box 1). This series is aimed mainly at policy-makers and provides a toolbox with many interesting and promising policy options. Most of the reports are presented in a hearing for members of parliament.

This chapter provides a short overview how CPB approaches the issue of ‘better data, for better studies for better policies’ using microdata. After having discussed CPB’s main type of activities in this area, Sect. 2 will briefly stipulate CPB’s position in the debate between the use of experimental methods and structural models. The chapter also provides some examples of recent research. It concludes with a summary of the challenges that must be faced while undertaking this type of research and how to overcome them.

2 Policy Evaluations with Microdata at CPB

2.1 The Pros and Cons of Policy Evaluation Studies and Structural Models

There are two major strands of literature using large micro datasets. These are policy evaluation studies and structural models. Both approaches have their strengths and weaknesses. The main issues are briefly summarized here.

[1] ‘Promising Policies’ (*Kansrijk beleid* in Dutch) is a series produced by CPB and PBL Netherlands Environmental Assessment Agency and SCP Netherlands Institute for Social Research. The publications are in Dutch.

Policy evaluation studies use randomized control experiments or ‘natural experiments’ (typically a policy intervention) as exogenous variation to estimate the impact of a particular policy (for an overview, see Angrist and Pischke 2008; Imbens and Wooldridge 2009). The basic idea is that the policy intervention cannot be influenced by the agents, which generates exogenous variation between the treated group and a control group. Popular techniques of studying policy evaluations are the use of instrumental variables, the differences-in-differences approach and regression discontinuity (Angrist and Pischke 2010). The strengths of these approaches are the identification of the causal impact of an intervention, the possibilities for replication and the relative ease of undertaking robustness analysis (Angrist and Pischke 2010). However, policy evaluation studies also have their weak points. One of these is the external validity of the effects on other environments and other target groups. Moreover, there is often a weak link between economic theory and the estimated effects. The latter point complicates interpretation of the results and often precludes welfare analysis (Heckman 2010; Keane 2010).

Structural models derive the estimating equations from economic theory and estimate policy invariant parameters, such as preference parameters. External validity is one strength of structural models. Another is the close link between theory and the estimated effects (Keane 2010). However, weak points are the identification of the causal relations, that the identifying variation is not always clear or credible and that replication and robustness checks are rather labour intensive (Angrist and Pischke 2010; Heckman, 2010).

Considering the strengths and weaknesses of both approaches, they appear to be the mirror image of each other. Hence, a fruitful way forward seems to combine the best of both worlds (Chetty 2009; Blundell 2010; Heckman 2010). In particular, in the ‘Promising Policies’ series (CPB 2015, 2016b), CPB uses the results of both strands of literature.

There are also two recent empirical approaches that formally integrate both strands of literature, starting from different strands. First, the so-called sufficient statistics approach uses economic theory to provide a theoretical underpinning of the treatment effect in policy evaluation studies and considers counterfactual policy reforms in the ‘vicinity’ of the policy reform on which the effect is estimated (Chetty 2009). The second approach is to ‘validate’ structural model with (natural) experiments (Todd and Wolpin 2006). Specifically, in this approach the authors estimate a structural model and compare the simulated effects of a policy reform with the treatment effect using policy evaluation methods.

In the following sections, this chapter considers four case studies using large micro datasets at CPB. In the first case, the authors validate a structural model for the labour participation of parents with young children, using the policy evaluation method of differences in differences. In the second case, the authors estimate the effect of tax rates on tax evasion by owners of small corporations, using discontinuities in the tax system. In the third case, the authors estimate the effect of teacher experience on children’s outcomes, using the random assignment of twins to different teachers. In the fourth case, the authors evaluate the introduction of performance-based payment schemes in the mental health-care sector.

2.2 *The Impact of Subsidies for Working Parents with Young Children on Labour Supply*

To promote the labour participation of parents with young children, governments employ a number of fiscal instruments. In the Netherlands, working parents with a youngest child up to 12 years of age receive a subsidy per hour of formal childcare and in-work benefits. Since only working parents qualify for these subsidies and benefits, they should promote labour participation of parents with young children. Unfortunately, it is largely unknown which policy works best for employment. Therefore, CPB studies the effectiveness of different fiscal stimuli in a structural empirical model of household labour supply and childcare use and validates this structural model with a policy evaluation study on a large reform in the period 2005–2009.

Bettendorf et al. (2015) use the differences-in-differences approach to estimate the effect of the 2005–2009 reform in childcare subsidies and in-work benefits for working parents. Indeed, over the period 2005–2009, there was a major increase in the childcare subsidy per hour, which halved the effective price for parents, and also a major increase in the in-work benefits for parents with a young child. In total, over the period 2004–2009, childcare expenditure increased from EUR 1 billion to EUR 3 billion, and expenditure on in-work benefits increased from EUR 0.4 billion to EUR 1.4 billion. The treatment group consists of parents with a youngest child up to 12 years of age. The control group consists of parents with a youngest child 12–17 years of age. By comparing the labour market outcomes for the treatment and control after the reform with those before the reform, Bettendorf et al. (2015) can isolate the effect of the policy intervention. They use data from the Labour Force Survey (*Enquete Beroepsbevolking*), from Statistics Netherlands, for the period 1995–2009.

Figure 1 gives a plot of the participation rate of women in the two groups, before and after the reform. The treatment and control group move (more or less) in tandem up to the reform. Considering the control group of women with a young child aged 12–17 years, Bettendorf et al. (2015) observe some convergence between the treatment group and this control group after the reform. Indeed, a regression analysis confirms that the reform had a statistically significant positive effect (an increase of 30,000 persons) on the participation rate of mothers with a young child up to 12 years old. Furthermore, Fig. 1 also shows the participation rate of a potential alternative control group, women without children. Women without children have a trend in the participation rate different from that of the treatment group and are therefore not a valid control group.

De Boer et al. (2015) use a structural model to estimate the preferences over income, leisure and childcare. They use a discrete choice model, which has become the workhorse of structural labour supply modelling (Bargain et al. 2014), and a large and rich administrative panel dataset for the Netherlands, the Labour Market Panel of Statistics Netherlands. The full dataset contains panel data for 1.2 million individuals for the period 1999–2009. The dataset combines information on income (from various sources, e.g. labour income, profit income and various types of benefits) and working hours from the Social Statistical File (*Sociaal Statistisch Bestand*), data from the municipalities (*Gemeentelijke Basisadministratie*) on demographic characteristics (e.g. gender, age, ethnicity, ages of children, household type), data from the Labour Force Survey (*Enquete Beroepsbevolking*) on the highest completed level of education and data on the price and use of childcare per child (*Kinderopvangtoeslag*). Thanks to the size of the dataset, the authors can estimate preferences for a large number of subgroups. Furthermore, they can account for a large number of observable characteristics and look at a large number of outcomes. The childcare information is available for the period 2006–2009, which is the time span used in the estimations for the structural model. Large-scale reforms in childcare subsidies and in-work benefits during this period benefit the identification of the structural parameters.

De Boer et al. (2015) then simulate the 2005–2009 reform with the structural model and compare the simulated results with the estimated effects of the policy evaluation study for validation. The results are reported in Table 1. The top of the table gives the results for the participation rate and hours worked of mothers and fathers with a youngest child up to 3 years (pre-primary school age). The bottom of the table gives the results for the participation rate and hours worked of mothers and fathers with a youngest child aged 4 to 11 years (primary school age). Table 1 shows that the results for the structural model are in line with the results of the policy evaluation study for mothers. The estimated effect on the participation rate of fathers is again much in line with the prediction from the structural model. For the intensive margin, for fathers with a young child of primary school age, the policy evaluation study suggests a smaller negative effect on hours worked per

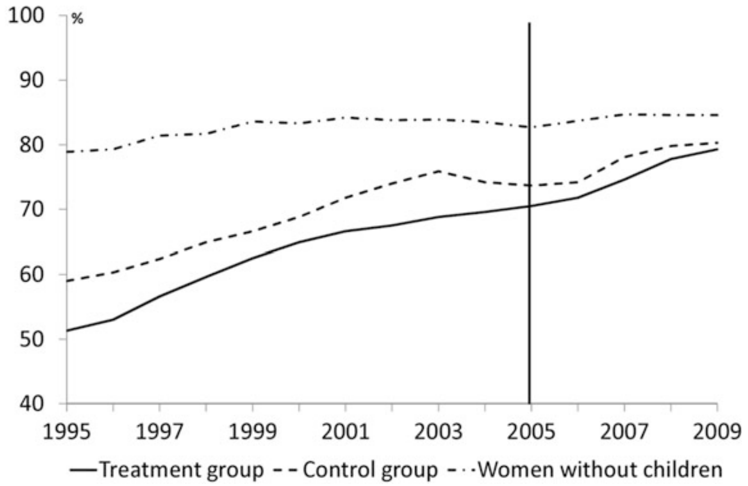


Fig. 1 Labour participation treatment and control group (differences-in-differences analysis). Source*:* Bettendorf et al. (2015)

Table 1 Comparison prediction structural model with policy evaluation study (differences in differences) for the 2005–2009 reform

Outcome	Structural model	Policy evaluation study (standard error in brackets)
Youngest child 0–3 years old (pre-school)		
Participation rate (mothers)	0.030	0.020 (0.007)
Hours worked per week (mothers)	1.185	1.222 (0.223)
Participation rate (fathers)	0.004	0.006 (0.004)
Hours worked per week (fathers)	0.075	−0.509 (0.237)
Youngest child 4–11 year old (primary school)		
Participation rate (mothers)	0.017	0.022 (0.007)
Hours worked per week (mothers)	0.616	0.750 (0.221)
Participation rate (fathers)	0.001	0.003 (0.004)
Hours worked per week (fathers)	−0.001	−0.180 (0.234)

Source*:* De Boer et al. (2015)

week than the structural model does, although the coefficients are not significantly different from each other. The only coefficient of the policy evaluation study which differs significantly from the prediction of the structural model is the hours worked response by fathers with a youngest child of pre-primary school age, for which the policy evaluation study suggests a larger negative response than the structural model.

Given that the structural model gives a good prediction of the estimated effect of the policy evaluation study, CPB uses this model to study the effectiveness of a number of counterfactual policy reforms for parents with young children. De Boer et al. (2015) find that an in-work benefit for secondary earners that increases with income is the most effective way to stimulate total hours worked. Childcare subsidies are less effective, as substitution of other nonsubsidized types of care for formal care drives up public expenditures. In-work benefits that target both primary and secondary earners are much less effective, because primary earners are rather unresponsive to financial incentives.

2.3 Tax Shifting by Owners of Small Corporations

In the Netherlands, owners of small corporations determine their own salary and the distribution of profits from their firm. This gives the owners the possibility to shift between tax bases. These owners are managing directors, abbreviated as DGAs (*directeur-grootaandeelhouder* in Dutch). DGAs thus face various types of taxation, such as taxation of corporate income, progressive taxation of labour income and proportional taxation of dividend income. As a consequence, they can exploit several opportunities to minimize the tax burden by shifting income between fiscal partners, between labour and dividend income and, in particular, over time.

From studies in other countries, it is known that the self-employed, like DGAs, are better able to avoid an increase in tax rates, because they face fewer frictions in shifting income to forms that are taxed at a lower rate (le Maire and Schjerning 2013; Devereux et al. 2014; Harju and Matikka 2016).

Tax shifting can be observed only using individual tax data of the DGAs and their corporations. Bettendorf et al. (2017) use the individual tax records of DGAs and the firms they own for the years 2007 until 2011 from the Tax and Customs Administration Office. There are about 300,000 DGAs per year and these DGAs own about 200,000 firms. Over time, their numbers are increasing.

Labour and capital incomes are taxed differently according to various boxes (Cnossen and Bovenberg 2001). Labour income is taxed at a progressive rate in the first box. Four tax brackets apply, ranging from 33% for incomes up to EUR 18,000 to the top marginal rate of 52% for incomes beyond EUR 56,000 in 2011. The salary of the DGA is also taxed in the first box. Different rules govern this salary. The rules state that the salary should be at least 70% of what is 'commonly' paid to managing directors of similar companies, the so-called reference salary. This reference salary also has a minimum level (EUR 41,000 in 2010). Many DGAs seem to consider this level an absolute minimum, whereas the correct interpretation is that the burden of proof shifts from the tax authority to the DGA at this level.

Distributed profits of DGAs are taxed in the second box. The profits from the sale of the company, or part of its shares, are also taxed in this box at a rate of 25%. The corporate tax rate is 20% for profits up to EUR 200,000 and 25% for profits exceeding EUR 200,000. This implies that the combined corporate and dividend tax rate is typically 40–44%. This hardly differs from the 42% tariff in the second and third tax bracket of the personal tax rate. For higher incomes facing the 52% tariff in the highest tax bracket, it then becomes attractive to shift income from wage income to profit income (see Fig. 2).

Bettendorf et al. (2017) find that taxable labour income, plotted in the panel on the right-hand side of Fig. 2, bunches at the cutoffs of the tax brackets, in particular at the top tax cutoff. The McCrary estimate of the discontinuity shows that the

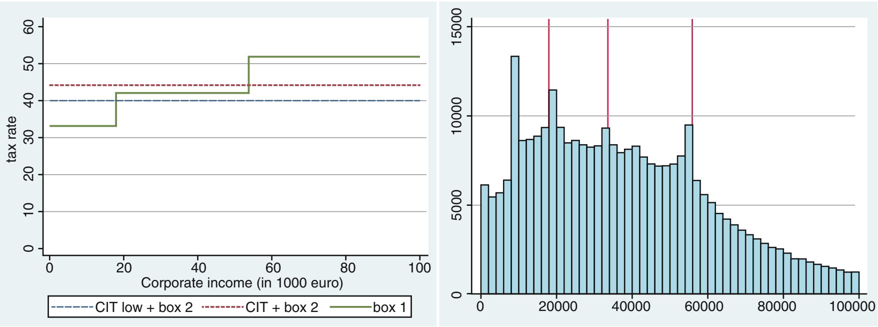


Fig. 2 Tax rate structure and distribution of wage income in the Netherlands, 2010. Source: Bettendorf et al. (2017)

density of gross wage income peaks exactly at this cutoff for all years (McCrary 2008). The elasticity of taxable income ranges from 0.06 to 0.11. Bettendorf et al. (2017) show that bunching at the top tax bracket cutoff is mainly driven by shifting income over time and to a much lesser extent by shifting between wage and profit income within a year.

The modest peak around the labour income tax base at nearly EUR 33,000 in the right-hand side figure may be surprising because, for people below the age of 65, the tax rate increases by only 0.05% moving from the second to the third tax bracket. This tiny increase in the tax rate cannot explain any bunching. However, for those aged 65 and older, the tax rate increases from 17% to 42% between the second and the third tax bracket. Some further analysis shows that a part of this peak is indeed explained by DGAs aged 65 and older. With a peak at EUR 18,000, the first tax cutoff is of nearly the same size as the last cutoff. The excess amount is somewhat smaller for this kink in the tax system than for the kink at the start of the top tax rate.

Bettendorf et al. (2017) conclude that DGAs manipulate their taxable wage and profit incomes to minimize their tax burden. In particular, they avoid the highest tax bracket. The salaries bunch just before the start of the fourth tax bracket. The effect is statistically significant, but on the other hand the economic impact in terms of forgone tax receipts is modest. The strict rules on the salaries prevent many DGAs from fixing their salaries just below the highest tax brackets. Their opportunities to determine their salaries are limited. These rules seem to achieve their goal, at least at the lower end of the wage distribution. The administrative burden is high, however, and discussions between the tax authorities and the DGAs consume time and effort. Moreover, DGAs seem to retain their profits in the firm instead of distributing them. Policy-makers seem to be aware of this situation. One of the loopholes, avoiding paying tax on dividends by emigration, has recently been closed. Other policy proposals are formulated to stimulate DGAs to distribute their profits more evenly over time, by introducing a two-rate tax structure in the second box of the personal income tax.

2.4 *Teacher Quality and Student Achievement: Evidence from a Sample of Dutch Twins*

The quality of teachers is considered to be a crucial factor for the production of human capital. Understanding the determinants of teacher quality is important for improving the quality of education and therefore a key issue for educational policy. A large literature has investigated the contribution of teachers to educational achievements of students (Hanushek and Rivkin 2006; Staiger and Rockoff 2010). A consistent finding in the literature is that teachers are important for student performance and that there are large differences between teachers in their impacts on achievement. However, the factors that are important for teacher quality remain unclear. The international literature suggests that the only factor that matters is teacher experience, but evidence is scarce and there are no results for the Netherlands. Gerritsen et al. (2016) investigate the extent to which this result also holds for the Dutch context.

Addressing this question is notoriously difficult because students, teachers and resources are almost never randomly allocated between schools and classrooms. Using nonexperimental methods may therefore yield biased results. For instance, more highly educated parents may select better schools or classrooms because they may be more involved with their children than less educated parents (Clotfelter et al. 2006; Feng 2009).

Gerritsen et al. (2016) try to circumvent these selection issues by examining the effect of teacher quality on student achievement using an experimental method. They use a novel identification strategy that exploits data on pairs of twins who entered the same school but were allocated to different classrooms in an exogenous way. The variation in classroom conditions to which the twins are exposed can be considered exogenous if the assignment of twins to different classes is as good as random. In many Dutch schools, twins are assigned to different classes because an (informal) policy rule dictates that twins are not allowed to attend the same class. As a result, they go to different classrooms. Because twins are more similar than different in early childhood, it seems unlikely that small differences between twins will affect the way they are assigned to different classes. In the empirical analysis, Gerritsen et al. (2016) have tested this assumption and did not find evidence for non-randomness of the assignment.

The research is designed to study classroom quality, as twins go to different classrooms. Classroom quality is a multidimensional concept that includes factors such as peer quality, class size and teacher quality. In the empirical analysis, Gerritsen et al. (2016) focus on the effects of observed teacher characteristics on student outcomes because, in applying this design, teachers seem the most obvious factor differing across classes (Dutch schools equalize other factors such as classroom facilities and class composition across classes).

For the analyses, longitudinal data of a large representative sample of students from Dutch primary education are used. The twins are identified from the population-based sample by using information on their date of birth, family name and school from the biannual PRIMA project. This project consists of a panel of approximately 60,000 pupils in 600 schools. Participation of schools in the project is voluntary. The main sample, which includes approximately 420 schools, is representative of the Dutch student population in primary education. An additional sample includes 180 schools for the oversampling of pupils with a lower socioeconomic background (the low-SES sample). Gerritsen et al. (2016) use all six waves of the PRIMA survey, including data on pupils, parents, teachers and schools from the school years 1994–1995, 1996–1997, 1998–1999, 2000–2001, 2002–2003 and 2004–2005. Within each school, pupils in grades 2, 4, 6 and 8 (average age, 6, 8, 10 and 12 years) are tested in reading and math. The scores on these tests are the main dependent variables. Information on teachers and classrooms consists of variables such as class size and teacher experience measured in number of years working in primary education. These are the explanatory variables.

In line with earlier studies on teacher effects, Gerritsen et al. (2016) find that teacher experience is the only observed teacher characteristic that matters for student performance (Staiger and Rockoff 2010; Chetty et al. 2011; Hanushek 2011). Twins that are assigned to classes with more experienced teachers perform better in reading and math. On average, one extra year of experience raises test scores by approximately 1% of a standard deviation. In the Dutch context, this means that at the end of primary education, a pupil taught by a teacher with 40 years of experience starts on average nearly one track higher in secondary education than a pupil taught by a new teacher. The effects of teacher experience are most pronounced in kindergarten and early grades. Gerritsen et al. (2016) also find that teacher experience matters in later career stages. This finding contradicts 'the consensus in the literature' that only initial teacher experience (less than 3 years) matters (Staiger and Rockoff 2010). However, the findings of Gerritsen et al. (2016) are consistent with the results found by Krueger (1999) and Chetty et al. (2011), using data from the STAR experiment, in which students and teachers were randomly assigned to classes, and also with recent findings by Wiswall (2013) and Harris and Sass (2011).

2.5 Evaluation of Performance-Based Payment Schemes in Mental Health Care

In 2008, the Dutch government introduced performance-based payment schemes in Dutch curative mental health care. Since these payment schemes were new, and their impact unclear, the government decided to initially apply the performance-based payment scheme only for a small group of mental health-care providers: the self-employed. The self-employed perform about 10% of all mental health services, while large mental health institutions such as psychiatric hospitals or regional facilities for ambulatory care perform the majority. These large mental health institutions receive an annual budget, and their employees, including psychiatrists, psychologists and mental health-care nurses, receive a fixed salary.

The main idea of policy research by Douven et al. (2015) is to evaluate the performance-based payment schemes for self-employed mental health-care providers using the large mental health institutions as a control group. Douven et al. use administrative data from the Dutch health-care authority covering the years 2008 to 2010. They have information from 1.5 million treatment records, where each record describes the total treatment episode of a patient from start to end. Depending on the severity of the patient's symptoms, a total treatment episode can take between a few hours and a whole year. Each treatment record contains detailed information about patient, treatment and provider characteristics. The records describe all curative mental health-care treatments that occurred in the Netherlands between 2008 and 2010.

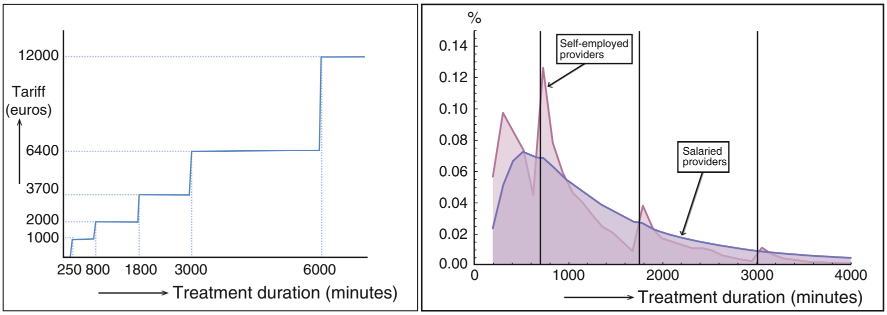


Fig. 3 Performance-based payment scheme and distribution of treatment durations. **(a)** Incremental payment function of self-employed providers (numbers on left axis are rounded off). **(b)** Treatment distribution of self-employed providers and salaried providers in large mental health institutions. Source*:* Douven et al. (2015)

A mental health-care provider diagnoses a patient and registers the severity of the patient's condition, the face-to-face treatment time with the patient, the type of treatment and daytime activities. The total time of a treatment is measured as a weighted sum of face-to-face time and daytime activities. The performance-based payment scheme depends on total treatment time and follows an incrementally increasing payment function. Figure 3a shows an example of the tariffs that providers receive for depression treatments. The incremental payment function jumps to a higher tariff when total treatment time passes a threshold of 250, 800, 3000 or 6000 min. Douven et al. (2015) distinguish between two types of financial incentives. First, there is an intended incentive. On the flat part of the payment scheme, a provider has no financial incentive at the margin to prolong treatment duration because his or her payment remains the same. Second, however, there is also an unintended incentive. A provider has an incentive to prolong treatment to obtain a higher financial reward. For example, at 2900 min a provider has a strong incentive to prolong treatment by 100 min to obtain a higher tariff. This almost doubles the income from that treatment, from EUR 3700 to EUR 6400.

Figure 3b shows the distribution of treatment durations for self-employed mental health-care providers and salaried providers in large mental health institutions. For the self-employed providers, bunching of treatment durations just after tariff thresholds is observed. For salaried providers, who do not get paid according to the performance-based payment scheme in Fig. 3, no bunching behaviour is visible.

Douven et al. (2015) measure both the intended and unintended incentive effects. For the intended effect, they find that treatment by self-employed providers is 2.6–5.6% shorter than that by salaried providers. However, the unintended effect of bunching around tariff thresholds is also present. Self-employed providers treat mental health-care patients 10–13% longer than salaried providers. Indeed, the unintended effect is stronger than the intended effect. Summing up all effects, Douven et al. find an increase in total costs of 2.5% to 5.3%, which is on average an increase of EUR 50–100 per treatment (where the average price of a treatment is about EUR 2000). Since self-employed providers treated about 236,000 patients during that period, total costs have increased by EUR 12–24 million.

This research has been useful because it provides a clear demonstration that health-care providers are sensitive to financial incentives. Accounting for the behavioural responses of health-care providers is therefore an important element in the design of a payment system. In the Dutch system of so-called regulated competition, health insurance companies will discipline health-care providers. However, until 2014 health insurance companies did not have information about the exact treatment duration of health-care providers. Thus, insurers had no opportunity to perform the type of analysis conducted in this paper. This is gradually changing; since 2014, health insurers have been able to obtain exact information about treatment durations and are also becoming more financially responsible for mental health-care cost containment.

3 Challenges and Solutions

Academics consider evaluations, experiments and analysis a *sine qua non* for progress in understanding the workings of the economy. However, policies are implemented in the political arena, where different mores apply. Differences in perspective help explain why the road to evidence-based policies is a bumpy one. Here the topic is addressed from three perspectives: policy questions, data and methodology and results and policy implications.

3.1 Policy Questions

Policies are not automatically subjected to evaluation. Politicians and civil servants are not always keen to have their policies evaluated. In general, the public reception of evaluations is asymmetric. Even in the best of times, not all policies will be successful. Failures will be showcased in the media, while good results tend to be ignored. In the short run, evaluations may resemble a game which you can only lose.

Eagerness to undertake experiments varies. Experiments provide a good way to obtain a sense of what works and what does not. The expectation of success is more circumscribed. This does make experiments more attractive. However, experiments take time, and politicians can feel pressured to skip the trial-and-error phase. Experiments are also criticized on moral grounds. If the policy works, it is deemed unethical to deny citizens access. The question is, of course, whether or not this is really the case. Is it ethically responsible to expose the public at large to policies that may prove to be ineffective and sometimes costly? In medical science, experiments are daily business, and the difference in outcomes between receiving or not receiving treatment can be (very) large. Ethical standards have been developed and ethical issues are assessed by specific boards.

There are various ways to enhance the chances of this type of analysis. The law can prescribe evaluations at regular intervals. In the Netherlands, all budgetary outlays need to be reviewed every 5 years. However, rules are no guarantee for success. A box-ticking exercise should be avoided; an evaluation of the process instead of the results is a shortcut, sometimes taken to avoid difficult questions. Educating civil servants can help to overcome reluctance for a proper evaluation of the effects of a policy. Alternatively, independent organizations can be established that are free to undertake this type of research. In the Netherlands, the Court of Auditors not only covers the classic audit questions but is also tasked with an assessment of the effectiveness of budgetary outlays. CPB is a government-funded independent organization that is free to undertake whatever research it deems fit. Good relations and familiarity with ministries through other activities (in CPB's case, the regular forecasts) can also help.

3.2 Data and Methodology

Microeconometric research lives or dies by good data. This is not a hurdle to be underestimated. Data are seldom cheap, access is not always easily ensured, and privacy issues need to be addressed. The approach taken here is a pragmatic one. Data are obtained mainly from the legislative bases where they are made available (e.g. Statistics Netherlands, the Dutch Health Care Authority). This is relatively cheap and helps to overcome privacy issues.

Sometimes, CPB tries to obtain data on a more ad hoc basis. On occasion, it buys or receives data from private companies (e.g. health insurers), which can be more expensive. Legal requirements to collect data in a proper manner, whenever a major policy proposal is implemented, would be a step forward for obtaining data.

The accessibility of microdata, and in particular of administrative data, has increased substantially in the past decade. More and more frequently, Statistics Netherlands uses administrative data instead of surveys. These data are made available via remote access by ministries and academics, for research projects. Because many datasets can be linked, this provides the researchers with a large and rich set of relevant variables, with sometimes millions of observations (or more).

Academic results of policy evaluations are often derived using sophisticated methodologies and have sometimes unexpected outcomes. In particular, if these results are not intuitively clear or desired by policy-makers, it is relatively easy to blame the method or the data. Sometimes, discussions with policy-makers help to explain methodologies and results more clearly, but this is not always the case. Moreover, discussions about the internal and external validity of results (see Sect. 2.1) are frequent. In policy evaluations, it is difficult to defend claims about the general validity of the results. This is not the case with structural models, but here other issues arise. It is often argued that a particular policy instrument is not well modelled or that it misses the specificities of the instrument, for example. This is often used as an argument that the outcomes of the model do not carry over to the policy instrument. In these discussions, it certainly helps if policy evaluations and structural models deliver the same results, as CPB's experience on the subsidies for working parents with young children show.

3.3 Results and Policy Implications

Negative results may lead to a hostile reception: 'killing the messenger' instead of dealing with the matter at hand. To make the reception of the results more effective, CPB lives by the rule 'no surprises'. Policy-makers may not like bad news, but they definitely hate surprises. A factual as opposed to a normative presentation of the results also helps.

Once evaluations or analyses have been undertaken, they are no guarantee that policies will be changed as a consequence. This frustrates many Dutch academics. What can be done? Naturally, results need to be readily available and easily accessible for policy-makers. This requires short notes, to the point and in layman's terms, with attractive infographics. More generally, a sound communication strategy will help. Moreover, persistence does pay off. The Dutch were at the forefront of analysing the consequences of ageing for public finances. Early on, it was evident that linking the retirement age to rising life expectancy would provide a strong antidote. While it took more than a decade to raise the official retirement age, when it was increased, the indexation followed shortly after. *Frappez, frappez toujours* is an important ingredient for success.

References

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

Angrist J, Pischke J-S (2010) The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. J Econ Perspect 24(2):3–30

Bargain O, Orsini K, Peichl A (2014) Comparing labor supply elasticities in Europe and the United States: new results. J Hum Resour 49(3):723–838

Bettendorf L, Jongen E, Muller P (2015) Childcare subsidies and labour supply: evidence from a large Dutch reform. Labour Econ 36:112–123

Bettendorf L, Lejour A, van 't Riet M (2017) Tax bunching by the owners of small corporations. De Economist 165:411–438

Blundell R (2010) Comments on: Michael P. Keane 'Structural vs. atheoretic approaches to econometrics'. J Econ 156:25–26

Chetty R (2009) Sufficient statistics for welfare analysis: a bridge between structural and reduced-form methods. Annu Rev Econ 1:451–488

Chetty R, Friedman JN, Hilger N et al (2011) How does your kindergarten classroom affect your earnings? Evidence from project STAR. Q J Econ 126(4):1593–1660

Clotfelter CT, Ladd HF, Vigdor JL (2006) Teacher–student matching and the assessment of teacher effectiveness. J Hum Resour 41(4):778–820

Cnossen S, Bovenberg L (2001) Fundamental tax reform in the Netherlands. Int Tax Public Financ 7:471–484

CPB (2015) Kansrijk Arbeidsmarktbeleid (Promising Labor Market Policies). CPB Netherlands Bureau for Economic Policy Analysis, The Hague

CPB (2016a) Kansrijk Arbeidsmarktbeleid 2 (Promising Labor Market Policies 2). CPB Netherlands Bureau for Economic Policy Analysis, The Hague

CPB (2016b) Kansrijk Onderwijsbeleid (Promising Education Policies). CPB Netherlands Bureau for Economic Policy Analysis, The Hague

De Boer H-W, Jongen E, Kabatek J (2015) The effectiveness of fiscal stimuli for working parents. IZA Discussion Paper No 9298. Institute for the Study of Labor, Bonn

Devereux M, Liu L, Loretz S (2014) The elasticity of corporate taxable income: new evidence from UK tax records. Am Econ J Econ Policy 6(2):19–53

Douven R, Remmerswaal M, Mosca I (2015) Unintended effects of reimbursement schedules in mental health care. J Health Econ 42:139–150

Feng L (2009) Opportunity wages, classroom characteristics, and teacher mobility. South Econ J 75:1165–1190

Gerritsen S, Plug E, Webbink D (2016) Teacher quality and student achievement: evidence from a sample of Dutch twins. J Appl Econ 32(3):643–660

Hanushek EA (2011) The economic value of higher teacher quality. Econ Educ Rev 30(2):466–479

Hanushek EA, Rivkin SG (2006) Teacher quality. In: Hanushek E, Welch F (eds) Handbook of economics of education, vol 2. Elsevier, Amsterdam, pp 1051–1078

Harju J, Matikka T (2016) The elasticity of taxable income and income-shifting: what is real and what is not? Int Tax Public Financ 23:640–669

Harris DN, Sass TR (2011) Teacher training, teacher quality and student achievement. J Public Econ 95:798–812

Heckman J (2010) Building bridges between structural and program evaluation approaches to evaluating policy. J Econ Lit 48(2):356–398

Imbens G, Wooldridge J (2009) Recent developments in the econometrics of program evaluation. J Econ Lit 47(1):5–86

Keane M (2010) A structural perspective on the experimentalist school. J Econ Perspect 24(2):47–58

Krueger AB (1999) Experimental estimates of education production functions. Q J Econ 114(2):497–532

le Maire D, Schjerning B (2013) Tax bunching, income shifting and self-employment. J Public Econ 107:1–18

McCrary J (2008) Manipulation of the running variable in the regression discontinuity design: a density test. J Econ 142(2):698–714

Staiger DO, Rockoff JE (2010) Searching for effective teachers with imperfect information. J Econ Perspect 24(3):97–117

Todd P, Wolpin K (2006) Using a social experiment to validate a dynamic behavioral model of child schooling and fertility: assessing the impact of a school subsidy program in Mexico. Am Econ Rev 96(5):1384–1417

Wiswall M (2013) The dynamics of teacher quality. J Public Econ 100:61–78

Rudy Douven is a health economist at CPB Netherlands Bureau for Economic Policy Analysis and the Erasmus University Rotterdam. He graduated in mathematics at the University of Eindhoven and received his PhD in econometrics from Tilburg University. During his career he obtained 1-year fellowships at Warwick University and Harvard Medical School. His main research interest is health economics.

Laura van Geest is Director of CPB Netherlands Bureau for Economic Policy Analysis. She graduated both in economics and public administration at Erasmus University Rotterdam. Before joining CPB, she worked at the Netherlands Ministry of Finance, most recently as Director General for the Budget and Treasurer General. Earlier in her career, she worked for the Dutch Constituency at the International Monetary Fund.

Sander Gerritsen is a researcher at CPB Netherlands Bureau for Economic Policy Analysis. He obtained his Master of Science in Econometrics from the University of Amsterdam and his PhD from Erasmus University Rotterdam. His main research interests are economics of education, labour economics and policy evaluation.

Egbert Jongen is a programme leader in labour markets at CPB Netherlands Bureau for Economic Policy Analysis and part-time associate professor at Leiden University. He obtained his MA from Maastricht University (cum laude) and his PhD from the Free University Amsterdam. His main research interests are the tax-benefit system and labour market policies. He recently headed the MICSIM project, the construction of a behavioural microsimulation model for the analysis of changes in the tax-benefit system in the Netherlands. A key element of the MICSIM project was the empirical analysis of microdata using both structural and quasi-experimental methods.

Arjan Lejour is a programme leader in tax policy and housing market at CPB Netherlands Bureau for Economic Policy Analysis and associate professor at Tilburg University. He obtained his MA from Erasmus University Rotterdam and his PhD from Tilburg University. His current research interests are international taxation of corporate profits and capital income, taxation of firms and self-employed, fiscal treatment of housing and the political economy of fiscal reforms.

Long-Term Effects of a Social Intervention for Vulnerable Youth: A Register-Based Study

Jacob Nielsen Arendt and Mette Verner

1 Introduction

The number of children with conduct disorder problems is a growing concern in many countries. Early life problems can settle into permanent life problems with huge individual and societal costs. Young people with conduct disorder problems are likely to perform poorly in school and often face difficulties pursuing further education or employment. They often face a risk of entering a life path involving criminal activity, drug abuse, and mental health problems. The evidence on effective interventions for disadvantaged adolescents is limited, as stressed by, for example, Nobel Laureate James Heckman and co-authors (Heckman et al. 2014). Some of these young people are placed outside their homes, in foster care, although little is known about the effectiveness of foster care (e.g., Doyle 2007; Frederiksen 2012).

Multisystemic therapy (MST) is an intervention developed in the United States (US) in the 1990s by psychologist Scott Henggeler and colleagues (e.g., Henggeler et al. 1992), with a focus on 12- to 17-year-old violent juvenile offenders. It is an intensive family-based intervention based on a systemic approach: the therapist meets the child in the family home or in other familiar surroundings and is available

This chapter was written while Jacob Arendt was employed at the Danish Institute for Local and Regional Government Research (now The Danish Centre of Applied Social Science).

J. N. Arendt (✉)
The ROCKWOOL Foundation Research Unit, Copenhagen, Denmark
e-mail: jar@rff.dk

M. Verner
The Danish Centre of Applied Social Science, Copenhagen, Denmark
e-mail: meve@vive.dk

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_14

on call 24 h a day, 7 days a week. MST has been evaluated in 55 published studies based on 25 randomized controlled trials and has been adopted in more than 34 US states and 15 countries worldwide (MST Services 2017). While this is a strong evidence base, meta-analysis shows great effect heterogeneity, especially across countries. MST has been implemented in Denmark since 2003, but to date there have only been two evaluations of the Danish program. The two evaluations applied a before–after design (Greve 2006; Greve and Thastum 2008), but no effect measurement was conducted.

Because the Danish implementers of the MST intervention have gathered social security numbers for all the participants, information from various public administrative registers can be linked by Statistics Denmark.

This chapter shows how the administrative register data can be used to construct effect measurements of MST in the Danish setting. It is based on a quasi-experimental effect design that utilizes access to data for the entire Danish youth population and their parents, from the time of the children's birth. By using such register data, some of the pitfalls in the existing studies can be avoided: most are based on small samples and have self-reported outcomes, measured over short time horizons. This creates a risk of self-evaluation bias and attrition bias, which can be avoided in an analysis based on administrative register data.

The chapter is organized as follows. In the next section, previous findings on the effect of MST are summarized. This is followed in Sect. 3 by a description of the data, and a description of the constructed control group is provided in Sect. 4. Section 5 presents the effect estimates. In Sect. 6 the findings are summarized, and the section concludes with a further discussion of the benefits of using register data for impact measurements.

2 Previous Literature

MST was developed in the late 1980s and 1990s by psychologist Scott Henggeler and colleagues at the Family Services Research Centre at University of South Carolina. More than 55 studies have been published on the effect and implementation of MST for different subpopulations of youth.

Many of these studies have been evaluated using a randomized controlled trial. Most of these have had a limited number of participants, from 16 (Borduin et al. 1990) to 176 (Borduin et al. 1995), and with Glisson et al. (2010) being an exception, with 615 observations.

Many of the studies have shown promising results on young people's behavior, family relations, and, for example, relapse into crime for juvenile offenders (see, e.g., Curtis et al. 2004). However, a Cochrane review from 2005 concluded that the evidence was not strong enough to make final recommendations for the use of MST in preference to alternative interventions (Littell et al. 2005). However, this review was based on only eight randomized controlled trials, and it should be stressed that no harmful effects of MST compared with alternative interventions were found and that the overall effect was positive, but not significantly different from zero. Three possible reasons for the difference in results when compared with previous reviews (e.g., Littell 2005) are highlighted in the Cochrane review: (1) previous reviews were narrative rather than systematic; (2) the Cochrane review excluded unpublished studies; and (3) the Cochrane review did not include studies not using the intention-to-treat principle. Two Dutch studies using administrative data point toward the aforementioned self-reporting bias as another potential explanation, finding positive effects for parent- and youth-reported problem behavior but not for criminal convictions (Asscher et al. 2013, 2014). Finally, there is a large amount of effect heterogeneity across countries and across different comparison groups. In some studies, MST is compared against another specific intervention, while in others it is compared against treatment as usual. This could in itself explain the divergent results.

The most recent review of the effects of MST includes 22 studies involving 4066 young people. This study finds that, after correcting for publication bias, there is no effect of MST on specific delinquency but a small, significant, and positive effect on general delinquency, psychopathology (e.g., externalizing and internalizing behavior), and family factors (e.g., self-reported family functioning, parenting skills, or parental health). The study confirms the finding of great effect heterogeneity, e.g., showing that the effect on delinquency is larger if MST is compared with a single alternative intervention (as opposed to multiple interventions) and if it is conducted in the United States. It also shows that, although study quality matters, the use of randomized versus quasi-experimental designs did not affect the results. The fact that the results of good quasi-experimental designs do not differ from those of randomized controlled trials has been reported elsewhere (e.g., Card et al. (2015) in the case of active labor market interventions), but of course this may differ from case to case (e.g., a famous study from economics by Lalonde (1986) and elaborations on his findings, e.g., by Dehejia and Wahba (1999)).

It is worth mentioning specifically the results from Norway and Sweden because the implementation setting in these countries is likely to most closely resemble the Danish implementation setting. Both were evaluated using a randomized controlled trial. While the Norwegian studies found that MST improves participants' internalizing symptoms and decreases out-of-home placements (Ogden and Halliday-Boykins 2004), the Swedish experience found no effect of MST when compared with standard alternative social interventions (Sundell et al. 2008).

The results from the reviews—reinforced by the divergent Scandinavian findings—show the need for case-by-case evaluation, which may well be conducted using quasi-experimental designs, at the very least as a cheap first-hand examination.

3 Data

The present analysis is based on information from Danish administrative register data. The relevant group of children were all observed from birth, and, furthermore, one can link information on parents to information on their children. The data come from various registers on education, health, crime, drug abuse treatment, labor market attachment, and the use of social measures. All the data were collected for administrative purposes, e.g., tax information in employment and health-care use, or for purposes of tax-financed reimbursement of health-care providers. The data are therefore considered to be highly reliable. As all Danish citizens have a unique identifier (CPR number), information from various public registers can be linked by Statistics Denmark.

The effect analysis is based on data on 436 participants who received MST at some point during the years 2007–2011. The participants were located in three municipalities (Herning, Aalborg, and Aarhus). At the time of entering MST treatment, the individuals were aged 12–17 years, and, since data are available up until 2013, the early participants can be followed up in the register data to the age of 22.

Table 1 shows the distribution of MST participation by age and the year of entering MST treatment. The majority of the participants entered the program around the age of 14–16 years, and there is a decline in the use of MST over the period.

Table 2 presents means of selected characteristics for MST participants and all other vulnerable young people aged 12–17 years in 2007–2011. "Other vulnerable young people" are defined as young people aged 12–17 receiving alternative social measures. The social measures under consideration are out-of-home placement (institution care or foster care) and preventive measures, such as a steady contact person. These measures are all offered with reference to the Danish Act on Social Services. The act implies both that vulnerable young people have a right to receive such measures and that local municipalities must pay for them.

Table 1 Distribution of entrance into MST

	Year											
Age	2007		2008		2009		2010		2011		All	
(years)	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%
12	5	4	7	9	5	6	4	5	0	0	21	5
13	14	12	10	12	12	13	9	10	7	12	52	12
14	22	19	18	22	13	14	22	25	11	19	86	20
15	30	26	21	26	19	21	18	20	17	29	105	24
16	30	26	16	20	26	29	20	23	13	22	105	24
17	16	14	10	12	15	17	15	17	11	19	67	15
All	117	100	82	100	90	100	88	100	59	100	436	100

Table 2 Background characteristics of MST participants and other vulnerable young people (no corrections)

	MST participants		Other vulnerable young people	
	Mean	Standard deviation	Mean	Standard deviation
Background characteristic, treatment year				
Age (years)	14.97	1.402	14.88	1.660
Girl	0.40	0.491	0.42	0.493
First- or second-generation immigrant	0.15	0.357	0.12	0.321
Living with single parent	0.61	0.487	0.72	0.448
Number of children in the family	1.94	1.159	1.17	1.317
Mother social welfare recipient	0.25	0.432	0.29	0.453
Father social welfare recipient	0.26	0.437	0.24	0.425
Characteristics year before treatment				
Father sentenced	0.13	0.335	0.12	0.319
Mother sentenced	0.03	0.182	0.05	0.223
Father in drug abuse treatment	0.01	0.117	0.02	0.154
Mother in drug abuse treatment	0.01	0.095	0.02	0.152
Father in psychiatric treatment	0.21	0.410	0.20	0.403
Mother in psychiatric treatment	0.37	0.117	0.37	0.482
Child in psychiatric treatment	0.23	0.421	0.24	0.425
Father prescription medicine	0.51	0.500	0.47	0.499
Mother prescription medicine	0.65	0.478	0.64	0.479
Child prescription medicine	0.30	0.457	0.31	0.462
Father hospitalized	0.37	0.583	0.36	0.479
Mother hospitalized	0.40	0.490	0.46	0.498
Child hospitalized	0.43	0.495	0.44	0.497
Institution care (1–2 years before)	0.05	0.219	0.26	0.437
Foster care (1–2 years before)	0.02	0.126	0.20	0.399
Other preventive measures (1–2 years before)	0.17	0.376	0.44	0.497
N	436		79,259	

The group means in Table 2 are given prior to any correction for background characteristics. The table shows that the two groups (MST participants and other vulnerable young people) are quite similar in terms of, for example, family background, parental characteristics, and health treatment. For instance, 13% of the MST participants have fathers with a criminal conviction. This is true for 12% of other vulnerable young people. It is also seen that a very large share of young people in both groups have parents who were in psychiatric treatment or hospitalized in the year before they received MST or an alternative social treatment. In contrast to these results, Table 2 also shows that the share of children and young people that were placed out of the home in the years before the comparison year was

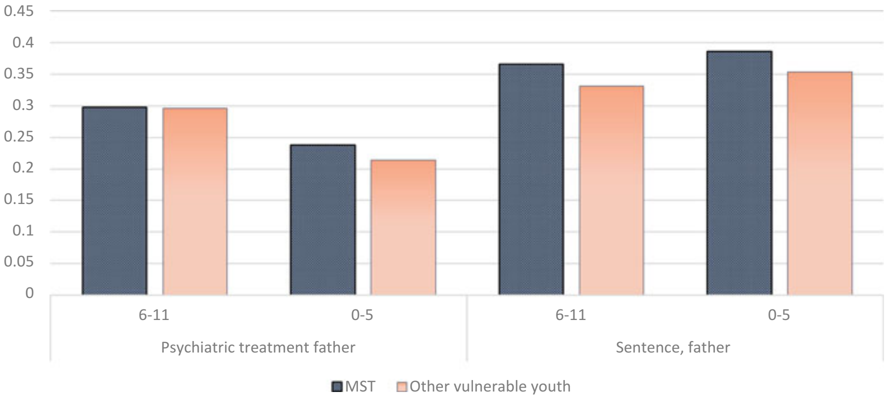


Fig. 1 Father's outcomes averaged over age groups for the child, for MST participants, and for other vulnerable young people

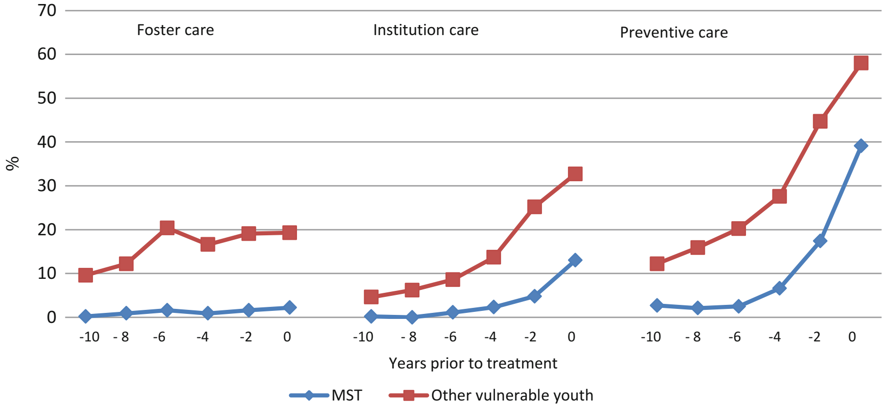


Fig. 2 Share of MST participants and other vulnerable young people receiving social measures in the years prior to treatment

substantially higher for the group of other vulnerable young people than for the MST participants. These patterns of similarity between the groups with respect to health-care usage and parental characteristics but divergence with respect to previous social interventions—are also found when looking further back in time during the children's lives. Figure 1 shows, as an example, the similarity in father's outcome for MST participants and other vulnerable young people, averaged over the years when the child was aged 0–5 years and when the child was aged 6–11 years. In contrast to this result, there is a marked difference between MST participants and other vulnerable young people with respect to the share who have previously received social treatment. This is shown in Fig. 2.

In a similar manner, the paths for some of the outcome variables of the analysis are compared in Fig. 3. Figure 3a–c shows the share of young people that was

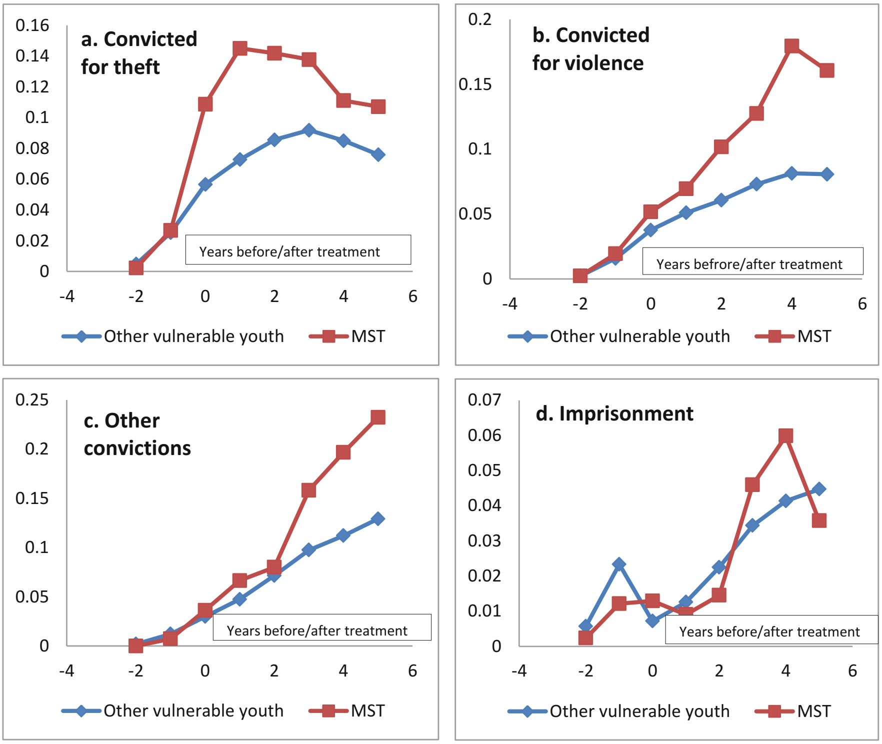


Fig. 3 **(a–d)** Share of young people convicted and imprisoned, by years before and after treatment

convicted in a given year. The conviction rates are shown both before and after the year of treatment. In the pretreatment period, the shares were close to zero for both groups, which is explained by the fact that the age of criminal responsibility in Denmark in the period of observation was 15 years. For all types of convictions, the rates increased after the treatment year, particularly for the MST participants. The gaps between MST participants and other vulnerable young people are as high as 10–15 percentage points. Hence, these descriptive measures do not indicate that MST has a crime prevention impact. It is, however, stressed that these numbers have not yet been corrected for differences in background characteristics between MST participants and other vulnerable young people and hence should not be interpreted causally. The corrected estimates are shown in the next section.

Figure 3d presents the share of young people sentenced to imprisonment. The difference between the two groups is lower and shifts over time. As imprisonment is a sentence reserved for more severe crimes, this indicates that the higher rates of convictions for MST participants can primarily be attributed to less severe types of crime.

4 Method

Evaluating an intervention such as MST is a challenge when the participants are not randomly selected from among the group of vulnerable young people. To handle the nonrandom selection, propensity score matching is applied to evaluate the effects of MST. This procedure matches each MST participant to a number of other vulnerable young people with a propensity to be treated with MST similar to that of the actual participants.

The MST participants were located in three municipalities (Herning, Aalborg, and Aarhus), and in this study, these participants are matched with similar young people in the same age group, who were living in other municipalities and receiving social measures. The matching is performed using the rich Danish register information on their childhood, health, family situation, and parents, observed every year from birth until the age at MST intervention.

As seen in the preceding section, the MST participants and other vulnerable young people are very similar, even without matching. What seems to be a major difference between the two groups is the extent to which they received social measures prior to the age at which they are compared. Prior social care is thus a key observable characteristic for which MST participants need to be matched with other vulnerable young people. Prior social care is also likely to be related to prior occurrence of conduct disorder and family problems. The design is illustrated in Fig. 4, highlighting that both the MST participants and their comparison groups are identified as vulnerable young people and therefore to a large extent similar at the outset. This study measures outcomes in the years after treatment and is able to follow MST participants and their comparison group for up to 5 years. Therefore, those who are MST participants at the age of 12 are tracked until the age of 17, and likewise those who are MST participants at the age of 17 are tracked until the age of 22. The follow-up period also depends upon the nature of the outcome, because education, crime, and employment outcomes are relevant at different ages.

The matching procedure includes register information on a very long list of variables including age, sex, family type, health-care history measures for children and their parents, criminal records, psychiatric treatment, drug abuse treatment, education of children's parents, and previous use of social measures.

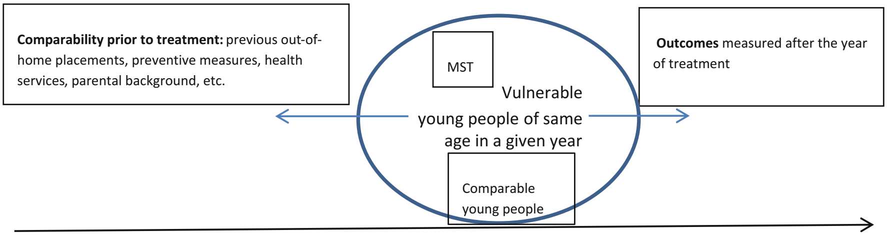

Comparability prior to treatment: previous out-of-home placements, preventive measures, health services, parental background, etc.
MST
Vulnerable young people of same age in a given year
Comparable young people
Outcomes measured after the year of treatment

Fig. 4 Evaluation design

Matching on this very rich set of characteristics of the child's past—including detailed annual lagged information on the social intervention history—enables a reliable comparison group to be identified for the effect evaluation (Lechner and Wunsch 2013).[1]

5 Effects

This study estimates effects of MST on outcomes measured from both short- and long-term perspectives. The outcomes include secondary education, youth education, employment, welfare dependency, criminal convictions, and use of social measures after MST treatment. All estimates are obtained by the use of matched controlled groups, as described in the previous section.

5.1 Education Outcomes

Compulsory schooling in Denmark is completed by a series of exams at the end of the 9th or 10th grade. The grade point average of the exam results is used as an outcome measure of secondary schooling, and the estimated effect (Table 3) is −0.083 and insignificant. However, a substantial share of vulnerable young people do not attend the exams, and, as is shown in Table 3, MST has a positive effect on the likelihood of attending the exam of 8.8 percentage points. It can be concluded, therefore, that although MST substantially increases the participation rate in exams, performance in the exams is not significantly affected.

From a longer-term perspective, it is relevant to see whether or not MST participants also complete the next level of education, youth education. Both branches of youth education, high school and vocational education, are taken into account. Table 4 presents the estimated effect of MST on the completion rate of youth education at various ages. For high school completion, no significant effects are found; whereas, for vocational education, the effect of MST is significantly negative at the ages of 20 and 22 years. This result is also seen in the overall completion rate of youth education, as the MST effect is −4.8 percentage points at the age of 20 and as high as −7.7 percentage points at the age of 22.

[1] Technically, propensity score matching with one nearest neighbor is used. The matched control group has good balancing properties, e.g., as seen in the relative difference on matched covariates between treated and controls after matching. It never exceeds 5%. The common support assumption is also fulfilled, and the estimates are not sensitive to the use of other matching techniques.

Table 3 Effect on secondary schooling outcomes

	Grade point average, mean (SE)	Attended exam, mean rate (SE)
Effect	−0.083 (0.122)	0.088*** (0.032)
Number treated	152	271
N	19,470	30,374

**p < 0.1, **p < 0.05, ***p < 0.01*

Notes: Estimates from matched control groups. SE = standard error

Table 4 Effect on youth education completion (high school and vocational education)

	Age (years), mean completion rates (SE)		
	20	21	22
High school education	−0.024 (0.013)	−0.008 (0.020)	0.010 (0.051)
Vocational education	−0.024 (0.018)	0.029 (0.026)	−0.087* (0.034)
Youth education, total	−0.048* (0.020)	0.021 (0.033)	−0.077* (0.015)
Number treated	105	71	39
N	15,709	10,715	6771

**p < 0.1, **p < 0.05, ***p < 0.01*

Notes: Estimates from matched control groups. SE = standard error

5.2 Employment and Welfare Dependency

The effect of MST on employment and welfare dependency is calculated at every age level between the ages of 18 and 22 years. It should be noted that the number of MST participants observed declines with age due to the distribution of MST participants as described in the previous section. Welfare dependency refers to any transfer income received from the state and includes, for example, social welfare payments, unemployment benefits, and study grants. Employment is registered as any month with positive earnings. Both welfare dependency and employment are registered as a number of weeks in a given year. Table 5 shows that, on average, duration of employment of MST participants is 2–2.5 weeks less than that of young people in the comparison group at ages 20–22. This also corresponds with a higher dependency on public transfer income, amounting to 2–3 weeks per year at ages 19–22. The two lower rows of estimates in Table 5 show the separate effect on weeks with study grants and weeks on social welfare. Study grants are universal and require only admittance to an ordinary education institution. The results show that the larger share of participants dependent on public transfers is mainly due to more time on social welfare and not more time in education with study grants.

Table 5 Effects on employment and welfare dependency

	Age (years), mean weeks (SE)				
	18	19	20	21	22
Employment	0.549 (0.845)	−0.527 (0.823)	−2.465* (1.019)	−2.048* (0.953)	−2.130** (0.816)
Welfare dependency	0.328 (0.911)	2.093* (1.021)	2.830 (1.132)	2.942* (1.231)	2.784* (1.009)
Study grants	−1.253 (0.841)	−0.163 (1.019)	2.747 (1.174)	0.282 (1.085)	3.901 (2.932)
Social welfare	2.203** (0.678)	2.802** (1.065)	1.023 (1.329)	3.935** (1.230)	1.909 (2.983)
Number treated	342	283	171	117	56
N	58,211	45,299	28,174	18,855	9252

$*p < 0.1$, $**p < 0.05$, $***p < 0.01$

Note: Estimates from matched control groups. SE = standard error

5.3 Crime

As seen in the descriptive section of this chapter, another highly relevant outcome in this evaluation is crime rate. Again, individuals are followed up to 5 years after the intervention year. Table 6 shows that most of the estimated effects, and in particular all significant effects, are positive. This means that MST participation *increases* the probability of having been convicted of a crime by up to 9.1 percentage points. No significant effects are found on the probability of being sentenced to imprisonment.

MST was originally developed as a measure directed toward criminal young people and has previously been shown to be effective in reducing young people's crime rates. Therefore, the sample is reduced to focus on young people who were convicted before the year of intervention, and similar effect estimates can be calculated for this subgroup. These are presented in Table 7. In this case, most effect estimates are insignificant. However, in the case of violence during the first year after the intervention, a significant negative effect of −5.5 percentage points is found, and, similarly, for imprisonment, the effect is −2.8 percentage points. This

Table 6 Effects on criminal convictions and prison sentencing

	Years after treatment, mean conviction rates (SE)				
	$t + 1$	$t + 2$	$t + 3$	$t + 4$	$t + 5$
Violence	−0.003 (0.014)	0.0287 (0.015)	0.047*** (0.009)	0.091*** (0.009)	0.075 (0.051)
Theft	0.057*** (0.016)	0.043* (0.019)	0.040* (0.018)	0.009 (0.024)	0.048 (0.042)
Other convictions	0.008 (0.013)	−0.002 (0.006)	0.038* (0.018)	0.049* (0.024)	0.104 (0.059)
Imprisonment	−0.003 (0.005)	−0.007 (0.007)	0.002 (0.013)	0.011 (0.014)	−0.005 (0.026)
Number treated	421	411	386	331	275
N	25,282	20,933	15,974	11,266	6835

$*p < 0.1$. $**p < 0.05$. $***p < 0.01$

Notes: Estimates from matched control groups. SE = standard error

Table 7 Criminal convictions and prison sentences, for individuals with previous convictions

	Years after treatment, mean conviction rates (SE)				
	$t + 1$	$t + 2$	$t + 3$	$t + 4$	$t + 5$
Violence	−0.055*** (0.234)	0.002 (0.015)	−0.005 (0.063)	−0.008 (0.077)	0.0346 (0.091)
Theft	0.018 (0.020)	0.021 (0.0422)	0.064 (0.059)	−0.077 (0.045)	0.004 (0.077)
Other convictions	−0.020 (0.172)	−0.014 (0.021)	0.009 (0.066)	−0.044 (0.087)	0.039 (0.101)
Imprisonment	−0.028** (0.131)	−0.006 (0.034)	−0.043 (0.046)	−0.031 (0.046)	−0.031 (0.064)
Number treated	71	65	56	39	26
N	3448	3224	2794	2312	1731

*p < 0.1. **p < 0.05. ***p < 0.01

Notes: Estimates from matched control groups. SE = standard error

suggests that, in this subsample of young people with previous criminal convictions, MST has reduced crime rates.

5.4 Social Measures

In a similar manner, the effect of MST on subsequent use of social interventions, in terms of out-of-home placement and preventive measures, can also be analyzed. From Table 8, it is seen that only in the first year after treatment is the number of days in out-of-home placement lower for the MST group than for the control group. However, caution should be taken when interpreting these effects. Because part of the control group is in out-of-home placement, which usually lasts more than a year, the control group will on average tend to have a higher out-of-home placement prevalence in the following year. For other preventive measures, the MST group receives significantly fewer days than the control group throughout the observation period, with the magnitude of the difference ranging between 15 and 50 days.

6 Discussion

This chapter has presented an example of how to use administrative register data to obtain new knowledge about the effectiveness of social interventions. It has examined the case of a small-scale social intervention—multisystemic treatment (MST)—which is targeted toward vulnerable young people and which has been

Table 8 Effects on out-of-home placements and preventive social measures

	Years after treatment, mean days (SE)				
	$t + 1$	$t + 2$	$t + 3$	$t + 4$	$t + 5$
Out-of-home placements	−36.260*** (7.285)	−6.695 (9.418)	−0.807 (10.009)	−9.988 (11.633)	13.909 (21.288)
Number treated	383	287	187	99	60
N	71,491	56,706	38,884	26,572	22,160
Other preventive measures	−50.7953*** (7.293)	−21.195*** (8.164)	−17.872** (8.881)	−15.524** (11.633)	−27.908*** (5.437)
Number treated	393	320	217	134	61
N	69,592	33,765	37,996	24,740	16,363

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: Estimates from matched control groups. SE = standard error applied to young people with conduct disorder problems and juvenile offenders. The intervention has been evaluated numerous times in various countries using randomized controlled trials but often with highly divergent results across target groups and across countries.

MST has been used in Denmark for more than 13 years, but its effect there has never been evaluated. The gold standard for effect evaluation is a randomized controlled trial. However, randomized controlled trials are expensive and are often difficult to conduct because of resistance from social workers and decision-makers when it comes to social interventions. But randomized controlled trials are not without other limitations, both with respect to potential contamination of control and treatment groups and because they are often based on small-scale interventions that are sometimes difficult to extrapolate to real-world settings. They are often also dependent upon survey measurements of outcomes, with problems of attrition over time and measurement error or self-reporting biases.

This is not to say that randomized controlled trials are not worth pursuing, but alternatives might be worth pursuing in difficult cases. Administrative register data are a hugely valuable source of information which, applied in the right way, can provide new insights at relatively low costs. This chapter has shown how to obtain effect measures, 13 years after the first use of MST in Denmark. The approach benefited from access to life-cycle data on children and their parents and the opportunity to track children and young people over time with very limited attrition and measurement error. The possibility that the estimates may have biases cannot be excluded, but it seems very unlikely that the biases are large enough to remove the negative effects documented.

References

Asscher JJ, Deković M, Manders WA et al (2013) A randomized controlled trial of the effectiveness of multisystemic therapy in the Netherlands: post-treatment changes and moderator effects. J Exp Criminol 9:169–187

Asscher JJ, Deković M, Manders WP et al (2014) Sustainability of the effects of multisystemic therapy for juvenile delinquents in the Netherlands: effects on delinquency and recidivism. J Exp Criminol 10(2):227–243

Borduin CM, Henggeler SW, Blaske DM et al (1990) Multisystemic treatment of adolescent sexual offenders. Int J Offend Ther Comp Criminol 35:105–114

Borduin CM, Mann BJ, Cone LT et al (1995) Multisystemic treatment of serious juvenile offenders: long-term prevention of criminality and violence. J Consult Clin Psychol 63:569–578

Card D, Kluve J, Weber A (2015) What works? A meta analysis of recent active labor market program evaluations. IZA Discussion Paper No 9236. Institute of Labor Economics, Bonn

Curtis NM, Ronan K, Borduin CM (2004) Multisystemic treatment: a meta-analysis of outcome studies. J Fam Psychol 18(3):411–419

Dehejia RH, Wahba S (1999) Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. J Am Stat Assoc 94(448):1053–1062

Doyle JJ (2007) Child protection and child outcomes: measuring the effects of foster care. Am Econ Rev 97(5):1583–1610

Frederiksen S (2012) Empirical essays on placements in outside home care. PhD thesis. In: Department of Economics and Business. Aarhus University, Aarhus

Glisson C, Schoenwald SK, Hemmelgarn A et al (2010) Randomized trial of MST and ARC in a two-level evidence-based treatment implementation strategy. J Consult Clin Psychol 78(4):537–550

Greve M (2006) Resultatevaluering af multisystemisk terapi i Danmark 2004–2007. Delrapport. Servicestyrelsen & Jysk Socialforsknings- og Evalueringssamarbejde, Højbjerg, p 1

Greve M, Thastum M (2008) Resultatevaluering af multisystemisk terapi i Danmark 2004–2007. Delrapport. Servicestyrelsen & Jysk Socialforsknings- og Evalueringssamarbej-de, Højbjerg, p 2

Heckman JJ, Humphrey JE, Kautz T (2014) The myth of achievement tests: the GED and the role of character in American life. University of Chicago Press, Chicago

Henggeler SW, Melton GB, Smith LA (1992) Family preservation using multisystemic therapy: an effective alternative to incarcerating serious juvenile offenders. J Consult Clin Psychol 60:953–961

Lalonde J (1986) Evaluating the econometric evaluations of training programs with experimental data. Am Econ Rev 76:604–620

Lechner M, Wunsch C (2013) Sensitivity of matching-based program evaluations to the availability of control variables. Labour Econ 21:111–121

Littell JH (2005) Lessons from a systematic review of effects of multisystemic therapy. Children and Youth Serv Rev 47:445–463

Littell JH, Popa M, Forsythe B (2005) Multisystemic therapy for social, emotional, and behavioral problems in youth aged 10–17. The Cochrane Database of Systematic Reviews 4:CD004797

MST Services (2017) Multisystemic therapy: research at a glance – published MST outcome, implementation and benchmarking studies. MST Services, Mount Pleasant

Ogden T, Halliday-Boykins CA (2004) Multisystemic treatment of antisocial adolescents in Norway: replication of clinical outcomes outside of the US. Child Adolesc Ment Health 9(2):77–83

Sundell K, Hansson K, Lofholm C et al (2008) The transportability of multisystemic therapy to Sweden: short-term results from a randomized trial of conduct-disordered youths. J Fam Psychol 22(4):550–560

Jacob Nielsen Arendt (PhD, Economics) is research leader at the Rockwool Foundation research unit. He was professor and program coordinator for Labor Market and Education at KORA from 2007 to 2011, and was associate professor in Health Economics at University of Southern Denmark from 2006 to 2011. His research focuses upon quasi-experimental impact evaluation and cost-benefit analysis of public interventions. He has expertise in the use of Danish administrative register data based on more than 20 years of use. Jacob has conducted research on social inequalities in health, the impact of financial aid for education, active labor market policy, and most recently on early child care interventions. He has published in well-renowned journals such as the *Journal of Health Economics*, *Empirical Economics*, and the *Journal of Human Capital* as well as more than 30 research reports and book chapters for Danish ministries and national boards.

Mette Verner is a research professor in economics at KORA, Danish Institute for Local and Regional Government Research, and she is also part of TrygFonden's Centre for Child Research at Aarhus University. Her work is focused on quantitative impact evaluations of policies and interventions directed toward families, children, and youth. Furthermore, within this area, Mette

Verner has specialized in cost-benefit analysis using survey and register-based data. In addition she has substantial experience within the field of empirical labor market and social research where her specialization is within the field of gender equality and family-friendly policies.

Does the Road to Happiness Depend on the Retirement Decision? Evidence from Italy

Claudio Deiana and Gianluca Mazzarella

1 Introduction

Retirement is a fundamental decision in the life-cycle of a person. For this reason many studies try to assess the effect of retirement on outcomes such as consumption, health, and well-being. Due to Italy's aging population, a trend which started in the second half of the twentieth century, it is fundamental to understand the effect of retirement on mental health and well-being. Charles (2004), among others, focuses on the effect of retirement in the United States (US), finding a positive effect of retirement on well-being. Using Canadian data, Gall et al. (1997) provide some empirical evidence in support of the theory first proposed by Atchley (1976), in which there is a positive short-term effect of retirement (defined as the *honeymoon* period) on well-being, but a negative mid- to long-term effect. Heller-Sahlgren (2017) identifies the short-term and longer-term effects of retirement on mental health across Europe. He shows that there is no effect in the short term, but that there is a strong negative effect in the long term. Bonsang and Klein (2012), Hershey and Henkens (2013), in Germany and in the Netherlands, respectively, try to disentangle the effects of *voluntary* retirement compared with those of *involuntary* retirement, finding that involuntary retirement has strong negative effects that are absent in the case of voluntary retirement. Börsch-Supan and Jürges (2006) analyze the German

Opinions expressed herein are those of the authors only. They do not necessarily reflect the views of, or involve any responsibility for, the institutions with which they are affiliated. Any errors are the fault of the authors.

C. Deiana · G. Mazzarella (✉)
European Commission, Joint Research Centre, Directorate I – Competences, Competence Centre on Microeconomic Evaluation (CC-ME), Ispra, VA, Italy
e-mail: claudio.deiana@ec.europa.eu; gianluca.mazzarella@ec.europa.eu

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_15

context and find a negative effect of early retirement on subjective well-being. Indeed, Coe and Lindeboom (2008) do not find any effect of early retirement on health in the US. The paper by Fonseca et al. (2015) shows a negative relationship between retirement and depression, but a positive relationship with life satisfaction. Finally, Bertoni and Brunello (2017) performed an analysis of the so-called *Retired Husband Syndrome* in Japan, finding a negative effect of the husband's retirement on the wife's health.

This chapter provides some new evidence of the effect of retirement on well-being, an effect which is characterized by self-reported satisfaction with the economic situation, health, family relationships, relationships with friends and leisure time, and by the probability of meeting friends at least once a week.

The remainder of the chapter is organized as follows: Sect. 2 illustrates the background of the literature on retirement. Section 3 details the data sources and provides some descriptive statistics. Section 4 illustrates the identification strategy and the empirical specification. Section 5 shows the effect of retirement on well-being, obtained using standard instrumental variables (IV) regression. Section 6 briefly discusses the Two-Sample Instrumental Variables estimator that will be applied to generalize result in Sect. 7. Conclusions follow.

2 Pension Reforms in Italy

Among developed economies, increasing life expectancy and reduced birth rates in the second half of the twentieth century have led to aging populations. In addition, empirical findings suggest an increase in anticipated retirement and consequently a reduction in the participation of elderly people at work (see, for example, Costa, 1998). These two trends have progressively unbalanced the ratio between retired and working people, compromising the financial sustainability of the social security system (see Galasso and Profeta, 2004).

This is primarily why policy-makers are typically deciding to increase the retirement age. Like many industrialized countries, Italy has experienced many pension reforms since the early 1990s. In Italy the first change in regulation was put in place in 1992, with the so-called Amato's law, which modified the eligibility criteria for the *old age pension*. Three years later, in 1995, a new regulation was introduced under the name of Dini's law. After a short period, the Italian government approved a new version (Prodi's law) in 1997. Finally, Maroni's law was implemented in 2004, which changed all of the eligibility criteria for the *seniority pension*.

This paper focuses on the changes made during the 1990s to the 'seniority pension' scheme. In particular, Dini's law introduced two alternative rules regulating pension eligibility, stating that the pension can be drawn either (1) at the age of 57, after 35 years of contributions, or (2) after 40 years of contributions regardless of age. As with Amato's law, the introduction of Dini's law was gradual, with age and contribution criteria increasing from 1996 to 2006, and with further evolution of

Table 1 Seniority pension: evolution of eligibility rules

Period	Age and contribution (years)	Contribution only (years)
–31/12/1995	–	35
01/01/1996–31/12/1997	52 and 35	36
01/01/1998–31/12/1998	54 and 35	36
01/01/1999–31/12/2000	55 and 35	37
01/01/2001–31/12/2001	56 and 35	37
01/01/2002–31/12/2003	57 and 35	37
01/01/2004–31/12/2005	57 and 35	38
01/01/2006–31/12/2007	57 and 35	39
01/01/2008–	57 and 35	40

the contribution criteria from 1996 to 2008. Prodi's law, in 1997, anticipated the changes of age and contribution criteria set by Dini's law. Table 1 summarizes the changes in the eligibility criteria provided by these laws.

The progressive tightening of the pension's requirements is associated with a decreasing retirement probability, which is evident comparing different cohorts given a certain age. However, neither law causes a *drastic* change in the retirement likelihood, and there is no expectation of a *discontinuity* at the threshold point, rather a gradual decrease provided by the progression of the law.

The individuals most likely to be affected by the reforms are those aged 52 so that we compare individuals at the same age but in different cohort. Table 1 summarizes the issue: before the reforms an individual was eligible to draw a pension after 35 years of contributions (having started work at 17, for example), but for the next 2 years would need to have 36 years' worth of contributions, and from 1999 would need 37 years' worth (which in the case of someone aged 52 would mean that they started work at 15, i.e. the minimum working age, and had no interruptions in their working career). Furthermore, workers cannot retire at any of the year because Dini's law also introduced the so-called *retirement windows*, fixed periods in which it is possible to stop working. For this reason most retirements are on 31 December and the first day of retirement is 1 January of the following year. So one would expect the first reduction in the number of retired workers to be in 1997. Due to differences in career paths, this study concentrates on male workers because females usually register more labor market interruptions to their working careers than men, and are automatically less affected by pension reforms.

3 Data

This section introduces the data sources used to obtain the two sets of results that are shown in Sects. 5 and 7. Furthermore, it provides some descriptive statistics on the variables considered.

3.1 Survey Data: AVQ

The study exploits a survey called *Aspetti della Vita Quotidiana* ((Aspects of Daily Life') (AVQ) carried out by the Italian Bureau of Statistics (Istat). It is an annual survey and each year involves about 50,000 individuals belonging to about 20,000 households, and it is a part of an integrated system of social surveys called *Indagine Multiscopo sulle Famiglie* ("Multipurpose Surveys on Household"). The first wave of the survey took place in 1993 and it includes different information about individual quality of life, satisfaction with living conditions, financial situation, area of residence, and the functioning of all public utilities and services.

All males aged 52 in the waves between 1993 and 2000 are selected, to give four cohorts from the pre-reform period and four from the post-reform period, for a total sample of 3143 individuals.

Table 2 presents the descriptive statistics of the outcomes involved in the analysis. Five outcome variables related to individual satisfaction were extracted from the AVQ, across various surveys. Respondents could choose from a Likert scale of four values, where a value of 1 means *Very much* and a value of 4 means *Not at all*. The authors created a set of dummy variables that are equal to 1 if the original variable is equal to 1, and 0 otherwise. A final dummy variable relates to the frequency with which individuals meet their friends, and takes a value equal to 1 if the answer is at least *once a week*. It is observed that almost 3% and 17% of the sample are satisfied with their economic situation and their health, respectively. More than 37% and 24% of the individuals are satisfied with their relationships with family and friends, respectively. The percentage of people who report satisfaction with leisure and meeting friends is 11% and about 70%, respectively.

3.2 Administrative Data: WHIP

The *Work Histories Italian Panel* (WHIP) is a statistical database constructed from administrative data from the National Institute of Social Security (INPS). It includes the work histories of private sector workers. INPS has extracted all the records

Table 2 Descriptive statistics

Sample	Variable	Mean	Standard error
AVQ	Satisfaction with the economic situation	0.029	(0.168)
	Satisfaction with health	0.164	(0.371)
	Satisfaction with family relationships	0.376	(0.485)
	Satisfaction with relationships with friends	0.241	(0.428)
	Satisfaction with leisure	0.114	(0.318)
	Meet friends at least once a week	0.690	(0.463)
	Retired	0.149	(0.356)

contained in its administrative archives relating to individuals born on 24 selected dates (irrespective of the year of birth), creating a sample size of about $1/15$ of the entire INPS population. The dataset is mainly structured in three different sections: the first relates to employment records, including yearly wage and type of contract; the second collects information on unemployment periods; and the third part is wholly dedicated to pensions, including first day of retirement, initial income from pension, etc.

The full sample initially included all male individuals aged 52 in the years covered by the AVQ survey, but the sample was not comparable with the survey data, mainly because the administrative data include individuals who cannot be included in the survey data (such as foreign citizens who have worked in Italy for just a few months, Italian citizens who have moved abroad and are therefore not represented in the survey data, etc.). For this reason, all individuals who worked less than 12 months in the 4 years between 1987 and 1990 were excluded from the sample (these years were selected to obtain a window that is removed from the years implemented in the analysis). The final sample includes 90,891 individuals.

4 Empirical Strategy

Retirement is a complex choice that involves multiple factors. This is an obvious reason why it is not possible simply to compare retired people with individuals who are not retired. These two groups are probably not comparable in terms of observed and unobserved characteristics. Indeed, one needs to look for an *exogenous variation* to help identify the effect of the retirement decision on well-being. In this context, this study exploits the changes to the pensions rules instigated by Dini's and Prodi's laws to instrument the retirement decision.

As summarized in Table 1, the progression provided by the two reforms does not allow identification of the retirement effects using a standard *regression discontinuity design* (see Hahn et al., 2001; Lee and Lemieux, 2010, for reviews). This is the reason why the effect of retirement is identified using the change of slope (kink) in the retirement probability. The identification strategy was first proposed by Dong (2016) and it mimics a binary treatment setting (where some individuals can be considered as treated and others as not treated) the *Regression Kink Design* (see Card et al., 2015; Nielsen et al., 2010; Simonsen et al., 2015). This allows the identification of the *local average response* for a continuous treatment setting (in which all the individuals can be considered as treated, but the amount of the treatment changes following certain predetermined rules). In this setting the change in slope at the threshold point becomes the additional instrument for the endogenous treatment decision (in this case the retirement choice). Then, the first-stage regression can be illustrated as follows:

$$D_i = \alpha_0 + \alpha_1(X_i - 1997) + \alpha_2(X_i - 1997)Z_i + \upsilon_i, \tag{1}$$

where D_i is a variable that is equal to 1 if the individual i is retired, 0 otherwise; X indicates the year in which the individual i reached age 52 and $Z = \mathbb{1}_{\{X \geq 0\}}$. The structural equation becomes:

$$Y_i = \beta_0 + \beta_1(X_i - 1997) + \beta_2 D_i + \epsilon_i, \tag{2}$$

where Y is the outcome of interest. The coefficient β_2 that comes from this specification corresponds to the ratio γ_2/α_2, where γ_2 is the coefficient related to $(X - 1997)Z$ in the *intention to treat* equation:

$$Y_i = \gamma_0 + \gamma_1(X_i - 1997) + \gamma_2(X_i - 1997)Z_i + \zeta_i,$$

and α_2 is as in Eq. (1) (see Appendix A in Mazzarella, 2015, for a formal proof). In this setting one can estimate Eq. (1) using both data sources, but the outcomes of interest Y are observed only in the survey data, so Eq. (2) can be computed only using AVQ data. The next sections present the results using the standard IV and TSIV estimators and then we compare the empirical evidence from the two. Specifically, we study the precision of the estimates born out from the survey and administrative data.

5 Results Using Survey Data: IV Estimates

This section discusses the main empirical results obtained using survey data. The first-stage coefficient (reported in the bottom row of Table 3) is equal to −0.0485

Table 3 Results

Panel	Outcomes	IV (AVQ)	TSIV (AVQ + WHIP)
(A)	Satisfaction	0.1094	0.1192
	with the economic situation	(0.1212)	(0.1259)
(B)	Satisfaction	0.0089	0.0098
	with health	(0.2576)	(0.2907)
(C)	Satisfaction	−0.0761	−0.0811
	with family relationships	(0.3488)	(0.3665)
(D)	Satisfaction	0.1086	0.1206
	with relationships with friends	(0.2953)	(0.332)
(E)	Satisfaction	0.4402*	0.4829**
	with leisure	(0.2254)	(0.2353)
(F)	Meet friends	0.6399*	0.7393**
	at least once a week	(0.3283)	(0.3555)
First stage	Coeff.	−0.0485***	−0.0419***
	Stand. err.	(0.0112)	(0.0022)
	Test F weak instrument	18.60	369.13

Standard errors in parentheses using bootstrap
$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

and is statistically significant at any level. This is consistent with the hypothesis that the reforms have progressively reduced the retirement probability. The F-statistic is equal to 18.60, which is larger than the threshold value of 10, so one can reject the hypothesis of weakness of the instrument.

The discussion now turns to the second-stage results. The first row in Table 3 shows the main findings. Each row shows different outcomes. The results in the first two rows demonstrate that retirement decisions are positively associated with an increase in economic and health satisfaction, even though statistical significance is not reached at any conventional level. On the one hand, one can observe a decrease in satisfaction with family relationships, but here too the estimates are not significant (third row). On the other hand, there is a positive relationship between retirement decision and satisfaction with relationships with friends, but again this is not significant.

That the decision to retire is generally associated with increased quality of relationships with friends can be related to better use of time—leisure versus work. In fact, the fifth row shows a positive relationship between retirement and satisfaction with leisure, and the sixth row reveals that retirement is associated with a higher probability of meeting friends at least once a week. Both coefficients are significant at 10%.

6 The Two-Sample Instrumental Variables Estimator

This section explains how the *two-sample instrumental variables* (TSIV) estimator can be implemented, since it is used to improve the precision of the estimates presented in Sect. 5.

The TSIV estimator was first proposed by Angrist and Krueger (1992) and, more recently, improved by Inoue and Solon (2010). It allows estimation of different moments from diverse data sources which are representative of the same population but which cannot be directly merged due to the lack of a unique identifier cross-database. The idea behind the TSIV estimator is to estimate the *first stage* regression using one sample, then use the coefficients estimated from this sample to compute the fitted value of the endogenous variables in the second sample. Finally, it exploits the fitted values to estimate the *structural equation* in the second sample.

Here this is briefly discussed in formal terms. $Y^{(s)}$ is defined as the $(n^{(s)} \times 1)$ outcome, $\mathcal{X}^{(s)}$ as the $(n^{(s)} \times p + k)$ matrix which includes the set of the endogenous (p) and exogenous variables (k), and lastly $\mathcal{Z}^{(s)}$ as the $(n^{(s)} \times q + k)$ (with $q \geq p$) matrix which comprises the set of additional instruments (q) and the exogenous variables, where $s = 1, 2$ denotes whether they belong to the first or to the second sample. The first-stage equations estimated with the second sample are, in matrix form:

$$\mathcal{X}^{(2)} = \boldsymbol{\alpha}\mathcal{Z}^{(2)} + \boldsymbol{v},$$

where $\boldsymbol{\upsilon}$ is an $(n^{(2)} \times p + k)$ matrix with the last k columns identically equal to zero. The previous equations could be estimated using standard ordinary least squares (OLS) to recover the value $\hat{\boldsymbol{\alpha}}$, which serves to obtain the fitted values of the endogenous variables in the first sample as:

$$\hat{\mathcal{X}}^{(1)} = \hat{\boldsymbol{\alpha}}\mathcal{Z}^{(1)}.$$

Finally, the structural equation could be estimated with the regression:

$$Y^{(1)} = \beta\hat{\mathcal{X}}^{(1)} + \epsilon.$$

The previous equations show how it is necessary to observe $(Y; \mathcal{Z})$ in the first sample and $(\mathcal{X}; \mathcal{Z})$ in the second sample, so $\mathcal{Z}$ has to be observed in both samples.

The TSIV estimator was originally proposed to allow the estimation of an IV regression when it is missing the required information of interest in both samples. In contrast, this study sheds some light on how the TSIV estimator can be used to improve the efficiency of the IV coefficient in estimating the first-stage regression with administrative data, even though the investigator can obtain the same information from survey data.

7 Results Combining Administrative and Survey Data: TSIV Estimates

This section presents results obtained by combining the survey data and the administrative data, in comparison with the standard IV results. The first-stage equation is estimated using WHIP, and the coefficients obtained with WHIP are then exploited to predict the fitted values of the endogenous retirement probability in AVQ. Finally the structural equation is estimated using AVQ. Standard errors are computed using the bootstrap method.

In general the estimates in Table 3 show how the TSIV estimator works with respect to the standard IV strategy. The first-stage coefficient is equal to -0.042 and it is highly statistically significant, with an associated F-statistic of 369.13. This is roughly 20 times larger than the coefficient obtained with survey data. The improvement of the precision of the first-stage estimates is also shown in Fig. 1, which compares the fitted values of the two samples, and their confidence intervals. All the sizes of the coefficients are almost unchanged, and they fall within the estimated confidence intervals of those calculated using survey data. The effects of retirement on satisfaction with economic situation, health, and relationships with family and friends are still not sizeable, and indeed the effects on satisfaction with leisure and on the probability of meeting friends at least once a week increase their significance from 10 to 5%, due to the increase of first-stage precision.

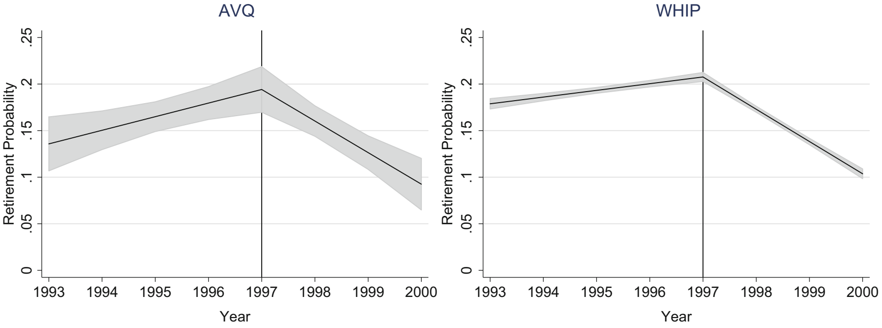


Fig. 1 First-stage comparison. (**a**) First stage with survey data. (**b**) First stage with administrative data

8 Conclusions

This study analyzes the retirement effect using as an exogenous variation the pension reforms that took place in Italy in the mid-1990s. It explains how to integrate survey and administrative data that are not directly linkable, estimating different moments from different data sources to obtain the *two-sample instrumental variables* estimator. In the empirical analysis all the required information is available in the survey data, but administrative data guaranteed a considerable improvement in the precision of the first-stage regression. The results from survey data are compared with those obtained by integrating the two data sources. The study shows that men increase their life satisfaction when they retire, providing further evidence that some men were adjusting their retirement decision, and that pension regulations prevented some men from locating precisely at the kink.

These results also have important implications. Administrative data have the advantage of giving detailed and precise information on large sample characteristics—in this case, retired men—over repeated periods of time. This chapter provides relevant evidence that the estimates' precision strongly depends on big data availability. This implies that policy-makers and politicians in general should foster access to administrative data to make the policy evaluation more systematic and estimates more accurate.

Appendix: Asymptotic Variance Comparison Using Delta Method

This appendix describes the conditions under which the estimator based on TSIV is more efficient than the simple IV estimator.

The approach is intentionally simplified and is based on delta method. Furthermore, it is assumed that the two samples are representative of the same population, and so the estimators are both unbiased for the parameter of interest.
The IV estimator β_{IV} could be defined as:

$$\beta_{\text{IV}} = \frac{\gamma_S}{\alpha_S},$$

where γ_S is the coefficient of the *intention to treat* regression and α_S the coefficient of the *first stage* regression, both computed using survey data (represented by the subscript S). They are asymptotically distributed as:

$$\begin{pmatrix} \hat{\gamma}_S \\ \hat{\alpha}_S \end{pmatrix} \dot{\sim} \mathcal{N}\left(\begin{pmatrix} \gamma \\ \alpha \end{pmatrix}; \begin{pmatrix} \frac{\sigma_\gamma^2}{n_S} & \\ \frac{\sigma_{\gamma,\alpha}}{n_S} & \frac{\sigma_\alpha^2}{n_S} \end{pmatrix}\right).$$

So the asymptotic variance of $\hat{\beta}_{\text{IV}}$ is equal to:

$$\mathbb{AV}\text{ar}\left[\hat{\beta}_{\text{IV}}\right] = \mathbb{AV}\text{ar}\left[\frac{\hat{\gamma}_S}{\hat{\alpha}_S}\right] = \frac{\sigma_\alpha^2/n_S}{\alpha^2} - \frac{\gamma^2\sigma_\gamma^2/n_S}{\alpha^4} - \frac{\gamma\sigma_{\gamma,\alpha}/n_S}{\alpha^3} \tag{3}$$

Similarly β_{TSIV} could be defined as γ_S/α_A (where the subscript A denotes the fact that it is computed with admin data), and:

$$\begin{pmatrix} \hat{\gamma}_S \\ \hat{\alpha}_A \end{pmatrix} \dot{\sim} \mathcal{N}\left(\begin{pmatrix} \gamma \\ \alpha \end{pmatrix}; \begin{pmatrix} \frac{\sigma_\gamma^2}{n_S} & \\ 0 & \frac{\sigma_\alpha^2}{n_A} \end{pmatrix}\right),$$

where the correlation between the two estimates is equal to 0 because they come from different samples.

Using similar arguments one can establish that the asymptotic variance of $\hat{\beta}_{\text{TSIV}}$ is equal to:

$$\mathbb{AV}\text{ar}\left[\hat{\beta}_{\text{TSIV}}\right] = \mathbb{AV}\text{ar}\left[\frac{\hat{\gamma}_S}{\hat{\alpha}_A}\right] = \frac{\sigma_\alpha^2/n_S}{\alpha^2} - \frac{\gamma^2\sigma_\gamma^2/n_S}{\alpha^4}. \tag{4}$$

From Eqs. (3) and (4) one can obtain that:

$$\mathbb{AV}\text{ar}\left[\hat{\beta}_{\text{TSIV}}\right] < \mathbb{AV}\text{ar}\left[\hat{\beta}_{\text{IV}}\right] \leftrightarrow \frac{\sigma_\alpha^2(n_A - n_S)}{\alpha^2 n_A} > 2\frac{\gamma\sigma_{\gamma,\alpha}}{\alpha^3}. \tag{5}$$

From Eq. (5) one can obtain the following conclusion:

1. If the policy has no effect (i.e., $\gamma = 0$), the TSIV estimator is even more efficient than the IV estimator (obviously if the sample size of the administrative data is bigger than that of survey data.)
2. Even if $n_A \to \infty$ it does not imply that $\mathbb{AV}\text{ar}\left[\hat{\beta}_{\text{TSIV}}\right] < \mathbb{AV}\text{ar}\left[\hat{\beta}_{\text{IV}}\right]$, and as a matter of fact Eq. (5) reduces to:

$$\frac{\sigma_\alpha^2}{\alpha^2} > 2\frac{\gamma\sigma_{\gamma,\alpha}}{\alpha^3},$$

   so the comparison still depends on quantities that could be both positive and negative (such as γ, α, $\sigma_{\gamma,\alpha}$).

References

Angrist JD, Krueger AB (1992) The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. J Am Stat Assoc 87(418):328–336

Atchley RC (1976) The sociology of retirement. Halsted Press, New York

Bertoni M, Brunello G (2017) Pappa ante portas: the effect of the husband's retirement on the wife's mental health in Japan. Soc Sci Med 175(2017):135–142

Bonsang E, Klein TJ (2012) Retirement and subjective well-being. J Econ Behav Organ 83(3): 311–329

Börsch-Supan A, Jürges H (2006) Early retirement, social security and well-being in Germany. NBER Technical Report No. 12303. National Bureau of Economic Research, Cambridge

Card D, Lee DS, Pei Z, Weber A (2015) Inference on causal effects in a generalized regression kink design. Econometrica 83(6):2453–2483

Charles KK (2004) Is retirement depressing?: Labor force inactivity and psychological well-being in later life. Res Labor Econ 23:269–299

Coe NB, Lindeboom M (2008) Does retirement kill you? Evidence from early retirement windows. IZA Discussion Papers No. 3817. Institute for the Study of Labor (IZA), Bonn

Costa DL (1998) The evolution of retirement. In: Costa DL (ed) The evolution of retirement: an American economic history, 1880–1990. University of Chicago Press, Chicago, pp 6–31

Dong Y (2016) Jump or kink? Regression probability jump and kink design for treatment effect evaluation. Technical report, Working paper, University of California, Irvine

Fonseca R, Kapteyn A, Lee J, Zamarro G (2015) Does retirement make you happy? A simultaneous equations approach. NBER Technical report No. 13641. National Bureau of Economic Research, Cambridge

Galasso V, Profeta P (2004) Lessons for an ageing society: the political sustainability of social security systems. Econ Policy 19(38):64–115

Gall TL, Evans DR, Howard J (1997) The retirement adjustment process: changes in the well-being of male retirees across time. J Gerontol B Psychol Sci Soc Sci 52(3):P110–P117

Hahn J, Todd P, Van der Klaauw W (2001) Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69(1):201–209

Heller-Sahlgren G (2017) Retirement blues. J Health Econ 54:66–78

Hershey DA, Henkens K (2013) Impact of different types of retirement transitions on perceived satisfaction with life. Gerontologist 54(2):232–244

Inoue A, Solon G (2010) Two-sample instrumental variables estimators. Rev Econ Stat 92(3): 557–561

Lee DS, Lemieux T (2010) Regression discontinuity designs in economics. J Econ Lit 48:281–355

Mazzarella G (2015) Combining jump and kink ratio estimators in regression discontinuity designs, with an application to the causal effect of retirement on well-being. PhD thesis, University of Padova

Nielsen HS, Sørensen T, Taber C (2010) Estimating the effect of student aid on college enrollment: evidence from a government grant policy reform. Am Econ J Econ Policy 2(2):185

Simonsen M, Skipper L, Skipper N (2015) Price sensitivity of demand for prescription drugs: exploiting a regression kink design. J Appl Econ 2(32):320–337

The Impact of Age of Entry on Academic Progression

Julio Cáceres-Delpiano and Eugenio P. Giolito

1 Introduction

This chapter studies the impact of age at entry on school performance using public administrative data for Chile.[1] In contrast to previous literature, the authors are able to track this impact on a selected group of outcomes by following a cohort of students over 11 years of their school life. This not only allows the authors to understand the evolution of the impact, but also sheds light on alternative channels that explain the pattern over time.

Since Deming and Dynarski (2008) pointed out a trend in the United States (US) towards delayed school entry, an increasing number of studies have explored

[1] Age at entry is expected to affect educational achievement over divergent channels, with various effects on individual outcomes. First, holding back a student is associated with a higher propensity to learn (school readiness). Second, delaying school entry means that students will be evaluated at an older age than other students who started earlier (age at test). In comparing children in the same grade, these two channels cannot be set apart from each other due to the perfect multicollinearity of these variables (Black et al. 2011). Third, by altering age at entry, parents affect the relative age of the students with ambiguous effect on educational outcomes. Fourth, the combination of minimum age entry rules and rules on mandatory education has been shown to affect dropout decisions (Angrist and Kruger 1991). Finally, a late start in school implies delaying entry into the labor market, that is, a reduction in the accumulation of labor experience (Black et al. 2011).

J. Cáceres-Delpiano (✉)
Universidad Carlos III de Madrid, Madrid, Spain
e-mail: jcaceres@eco.uc3m.es

E. P. Giolito
ILADES/Universidad Alberto Hurtado, Santiago, Chile

IZA, Bonn, Germany
e-mail: egiolito@uahurtado.cl

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_16

the short- and long-term effects of age at entry.[2,3] Despite the observed positive correlation between age at entry and academic achievements (Stipek 2002), a series of recent studies reveal mixed results for these and other long-term outcomes once the endogeneity of age at entry is addressed. Among children in primary school, age at entry has been negatively associated with grade retention and positively linked with scores in standardized tests (McEwan and Shapiro 2006). The evidence for older students shows that age at entry is associated with lower IQ and increased rates of teen pregnancy and mental health problems (Black et al. 2011). Earlier age at school entry is also associated with a negative effect on completed years of education (Angrist and Kruger 1991), lower earnings in young adulthood (20s and early 30s) (Angrist and Kruger 1991; Black et al. 2011), and an insignificant effect on the Armed Forces Qualifying Test (Cascio and Lewis 2006). Therefore, although a positive effect occurs in early school outcomes, age at entry seems to have an ambiguous long-term effect.

In this chapter, the analysis is focused on the impact of age at school entry on children of school age. This work is closely related to that of McEwan and Shapiro (2006), who also study the benefits of delaying enrollment on educational outcomes in Chile. They show that an increase of 1 year in the age of enrollment is associated with a reduction in grade retention, an increase in grade point average (GPA) during the first years, and an increase in higher education participation. Also, like McEwan and Shapiro (2006), the authors address the endogeneity of age at entry by using the quasi-random assignment stemming from the discontinuity in age at entry produced by the minimum age requirements. In the present study, however, the authors use the complete population of students.[4] This is important given the recognized sorting of better students into schools with better teachers (Tincani 2014). Second, the authors follow a cohort of students for over 11 years of their school life. By following this cohort, there is no need to restrict the analysis to a specific grade, which is a function of the number of grades repeated; rather, the analysis considers the number of years elapsed since a child started primary school. This is particularly relevant for Chile, where over 30% of the students in a particular cohort have repeated a grade at least once during the course of their school life. Third, by following a cohort of students for 11 years, the authors investigate the evolution of the impact of age at entry and its reported decline over the student's school life. In Chile, in contrast to other countries, laws on mandatory education are linked to years of education rather than to a student's age. In fact, for the period under analysis, the completion of secondary education is mandatory. This institutional feature enables study of

[2] As mentioned by the authors, one-fourth of this change is explained by legal changes, while the rest is attributed to families, teachers, and schools.

[3] A related part of the literature has focused on the impact of age per se, rather than age at entry. Among these studies is that of Kelly and Dhuey (2006), who found, among a sample of countries that are members of the Organisation for Economic Co-operation and Development (OECD), that younger children obtain considerably lower scores than older ones at fourth and eighth grades.

[4] McEwan and Shapiro (2006) focus part of their analysis on publicly funded schools in urban areas.

the long-term return of delaying school entry over a child's school life, even into secondary education, without concern that the impact of dropping out is captured simultaneously. Finally, by using other outcome variables going beyond school achievements, the authors are able to study the channels by which age at entry affects educational performance, and to assess the evolution of the impact associated with a delay in the age at entry. Specifically, the authors study the impact on school type, the type of academic track followed by the students and whether or not delaying school entry is associated with being in a school where higher *cream skimming* can be observed.[5]

The findings confirm that delaying school entry has a positive effect on GPA, attendance, and the likelihood of passing a grade, but also that this impact tends to wear off over time.[6] Nevertheless, in contrast to previous studies, the findings reveal that this impact may still be observed 11 years after a child has started school. Moreover, evidence on the effect of age at entry on school type provides a potential explanation for the decline in the impact of age entry on academic achievement over a child's school life. Specifically, a higher age at entry decreases the likelihood that a child is enrolled in a municipal school, such schools being characterized by less active selection of students and lower-quality teachers. Consistent with these differences in academic selection, children who are older when they start school have a higher probability of being enrolled in schools where children coming from other schools have a GPA that is higher than the mean in the school where the student was enrolled for first time. There is also evidence that age at entry has a positive effect on the likelihood that a child follows an academic track in high school. Finally, the authors provide evidence that age at entry is associated with an increase in the probability that a child is enrolled in a school which is actively engaged in *cream skimming*, which also explains the drop in the impact of age at entry on school achievements.

The chapter is organized as follows. Section 2 describes the authors' empirical strategy. Section 3 briefly sketches Chile's educational system, presents the dataset

[5] The term *cream skimming* has been used in the voucher literature to describe the fact that a voucher is more likely to be used by the parents of "high-ability" students or, more generally, students who are less costly to teach. These parents/students move from public, lower-performing, schools to better, private, ones, leaving the students who are more costly to teach to the public schools. In this chapter the concept of cream skimming is used to describe the active or passive process by which some schools end up capturing high-achievement students while other schools, as a result of this process, are left with students from lower levels of capacity/achievement.

[6] As mentioned above, (footnote 1), the channels by which age at entry can affect individual outcomes are diverse. Depending on which channel is more important, it will be more likely to observe a particular long-lasting effect associated with holding back a student. For example, it is argued that an *age at test* channel should wear off over time (Black et al. 2011). So, whether or not a positive long-lasting effect may be observed is an empirical question. Moreover, this study explores an additional channel that is often ignored, but which is of relevance in a system using a voucher scheme. This is the progressive sorting of students into schools, which we refer to in this chapter as *cream skimming*.

used in the analysis, and defines the sample and the selected outcomes in the analysis. In Sect. 4 the results are presented, and Sect. 5 concludes.

2 Empirical Specification

The specification of interest in this analysis can be expressed as follows:

$$y_{it} = \alpha_t + \gamma^t Aentry_i + \beta^t * X_i + g^n(x_i^n) + v_{it} \tag{1}$$

with y_{it} as one of the educational outcomes observed for a student, i, t years after she/he started school. $Aentry_i$ corresponds to the age at entry and X_i to other predetermined variables. The parameter of interest is γ^t, which corresponds to the impact of age at entry on a selected outcome. The time superscript highlights the fact that this impact is allowed to change over time. As extensively reported in the literature, it is suspected that estimating Eq. (1) by ordinary least squares (OLS) will produce inconsistent estimates of $\gamma.^t$ Families who decide to delay school entry are more likely to have relatively higher (lower) gains (costs) associated with this delay.[7] Unobserved variables correlated with these gains and costs, such as parents' education, parents' motivation, and so on, can also have a direct effect on a student's achievement; the OLS estimates are likely to pick up the impact of these unobserved factors as well as the impact of age at entry.

To overcome the problem of this omitted variable and to estimate the impact of age at entry, γ^t, the minimum age at entry rules is used as a source of variation in the age at enrollment in first grade of primary school. These rules establish that children, in order to be enrolled in first grade at primary school, must have turned 6 before a given date in the academic year. Children whose birthday occurs before this cutoff date are entitled to start school the year they turn 6. Those whose birthday is after this cutoff must wait until the next academic year to start school. This discontinuity in the age at entry, together with the assumption that parents cannot fully control the

[7] Parents decide to delay a child's school entry based on individual benefits and costs. On the side of benefits, the literature has stressed the concept of "readiness" for learning (Stipek 2002). On the cost side, families must face the cost of the labor income forgone by the household member responsible for childcare, or the market cost of childcare for the unenrolled children. These and other factors defining the decision to delay school entry are not always observed by the researcher, and are potentially correlated with school outcomes. For example, a mother's labor force participation has been shown to affect her child's health, which might affect school performance (Sandler 2011).

date of birth, provides a potential quasi-experimental variation in the age at entry that constitutes the core of our "fuzzy" regression discontinuity (RD) strategy.[8]

Chile's official enrollment cutoff used to be 1 April but, since 1992, the Ministry of Education has provided some degree of flexibility to schools, allowing them to set other cutoffs between 1 April and 1 July. In fact McEwan and Shapiro (2006) show that in practice four cutoff dates are used in Chile: 1 April, 1 May, 1 June, and 1 July.

Hahn et al. (2001) show that the estimation of causal effects in this regression discontinuity framework is numerically equivalent to an instrumental variable (IV) approach within a small interval around the discontinuity.[9] In particular, following the equivalence with an IV approach, the instruments in the analysis correspond to four dummy variables taking a value of 1 for those children whose birthday is just after one of the cutoffs, and 0 otherwise.

Finally, in Eq. (1), we include $g^n(x_i^n)$, which is a flexible polynomial specification in the date of birth for a student whose birthday is around the n cutoff.[10] By including $g^n(x_i^n)$ in the previous equation the authors recognize that students born at different times of year might differ in a systematic manner. In fact, Buckles and Hungerman (2013), for the US, show that season of birth is correlated with some mothers' characteristics. Specifically, they show that the mothers of children born in winter are more likely to have lower levels of education, to be teen mothers, and to be Afro-American. The fact that these mothers' characteristics are correlated simultaneously with birthday and child's educational outcomes does not invalidate the RD approach taken here. The fact that these observed and unobserved factors do not change discontinuously in the mentioned cutoffs is the basis for this identification.

[8] In a fuzzy RD treatment, status is not perfectly determined by the running variable. On the other hand, where the treatment is a deterministic function of date of birth, the probability of treatment would change from 1 to 0 at the cutoff day. For more details, see Lee and Lemieux (2010).

[9] By focusing on the observations around these four discontinuities, the study first concentrates on those observations where the age at entry is as if it were randomly assigned. This randomization of the treatment ensures that all other factors (observed and unobserved) determining a given outcome must be balanced on either side of these discontinuities. Second, and for a given parametrization of $g(.)$, the estimated function can be seen as the non-parametric approximation of the true relationship between a given outcome and the variable date of birth, that is, it is less likely that the estimated impacts are driven by an incorrect specification of $g^n(.)$. In particular, use is made of a bandwidth of 15 days around each of the discontinuities in the baseline specification. Two of the most popular methods are used to define this bandwidth. First, using the rule-of-thumb approach (Lee and Lemieux 2010), the bandwidth ranges from 5 to 10 days. Alternatively, using a method based on the calculation of the cross-validation function (Lee and Lemieux 2010) leads to a bandwidth of 20 days around the discontinuities. Moreover, the authors show in the working paper version that their results are robust to different bandwidths but also to the degree of $g(.)$.

[10] Specifically, $x_i^n = BD_i - C,^n$, where BD_i is the birth date in the calendar year of a student and C^n is one the four cutoffs.

3 Data and Variables

Since a major educational reform in the early 1980s,[11] Chile's primary and secondary educational system has been characterized by its decentralization and by a significant participation of the private sector. By 2012, the population of students was approximately 3.5 million, distributed in three types of school: public or municipal (41% of total enrollment), non-fee-charging private (51% of total enrollment), and fee-charging private (7% of total enrollment).[12] Municipal schools, as the name indicates, are managed by municipalities, while the other two types of schools are controlled by the private sector. Though both municipal and non-fee-charging private schools receive state funding through a voucher scheme, only the latter are usually called voucher schools.[13]

Primary education consists of 8 years of education while secondary education depends on the academic track chosen by a student. A "Scientific-Humanist" track lasts 4 years and prepares students for a college education. A "Technical-Professional" track in some cases lasts 5 years, with a vocational orientation aiming to help transition into the workforce after secondary education. Until 2003, compulsory education consisted of 8 years of primary education; however, a constitutional reform established free and compulsory secondary education for all Chilean inhabitants up to the age of 18. Despite mixed evidence on the impact of a series of reforms introduced in the early 1980s on the quality of education,[14] Chile's primary and secondary education systems are comparable, in terms of coverage, to any system in any developed country.

[11] The management of primary and secondary education was transferred to municipalities, payment scales and civil service status for teachers were abolished, and a voucher scheme was established as the funding mechanism for municipal and non-charging private schools. Municipal and non-fee-charging private schools received the same rates, tied strictly to attendance, and parents' choices were not restricted by residence. Although with the return to democracy some of the earlier reforms have been abolished or offset by new reforms and policies, the Chilean primary and secondary educational system is still considered one of the few examples in a developing country of a national voucher system. In 2009 it covered approximately 93% of primary and secondary enrollment. For more details, see Gauri and Vawda (2003).

[12] There is a fourth type of school, "corporations," which are vocational schools administered by firms or enterprises with a fixed budget from the state. In 2012, they accounted for less than 2% of the total enrollment. Throughout this analysis, they are treated as municipal schools.

[13] Public schools and subsidized private schools may charge tuition fees, with parents' agreement. However, these complementary tuition fees cannot exceed a limit established by law.

[14] The bulk of research has focused on the impact of the voucher funding reform on educational achievements. For example, Hsieh and Urquiola (2006) find no evidence that school choice improved average educational outcomes as measured by test scores, repetition rates and years of schooling. Moreover, they find evidence that the voucher reform was associated with an increase in sorting. Other papers have studied the impact of the extension of school days on children's outcomes (Berthelon and Kruger 2012), teacher incentives (Contreras and Rau 2012), and the role of information about the school's value added in school choice (Mizala and Urquiola 2013). For a review of these and other reforms since the early 1980s, see Contreras et al. (2005).

The data used in this analysis come primarily from public administrative records on educational achievement provided by the Ministry of Education of Chile for the period 2002–2012. These records contain individual information for the whole population of students during the years that a student stays in the system. Moreover, an individual's unique identification allows each student to be tracked over her/his whole school life. This dataset is essential to this study for four reasons. First, for every student in the system, there are several measures of school performance. Second, for the cohort of children born in 1996 or later, one can observe the age at which students were enrolled in first grade of elementary school. In fact, we define "age at entry" as the age at which a student is observed at the beginning of the school year when she/he was enrolled in first grade. The analysis is focused on the oldest cohort that was eligible to start school the first year for which there are data, that is, children born in 1996. These children, in complying with school's minimum age at entry rule, should have started school either in 2002 (those eligible to start school the year they turned 6) or in 2003 (in the case of those children whose entry into school was delayed until the year they turned 7). Third, these administrative records provide an opportunity to follow every child over her/his complete school life, and to analyze the impact of age at entry beyond the first year of enrollment. Depending on the age at entry in primary school, these children are observed either 11 or 12 times (years) in the records. Given this last constraint, the analysis is focused on the impact of age at entry over the first 11 years of a student's school life. Finally, because the whole population of students is observed, and the student's school can be identified at every year, it is possible to determine not only some characteristics of the school, but also whether or not a student is enrolled in a school engaged in *cream skimming*.

By using students' records two sets of outcomes are constructed. The first group of variables attempts to characterize the impact of age at entry on school performance. The first variable, attendance, corresponds to the percentage of days, out of the total, that a child has attended school during a given school year. Attendance, however, might well be capturing a school's effort, since, for those institutions receiving funding through the voucher system, funding is a function of students' attendance. To account for the fact that attendance is able to capture this and other school effects, we define a dummy variable "attendance below the median," which takes a value of 1 for students with attendance below the median in the class, and 0 otherwise. The next variable is the annual average GPA over all subjects. As well as the variable attendance, GPA could reflect a school's characteristics rather than a student's own achievements.[15] As with attendance, a dummy variable is defined as indicating whether or not the GPA for a student in a particular year is above the class median. Finally, for this group of outcomes, the variable "Pass" is defined as a dummy variable taking a value of 1 when a student passes to the next grade, and 0 otherwise.

[15] Anecdotal evidence exists on grade inflation, which has not been equally observed among all schools.

The second group of variables is composed of variables describing the movement of students between schools and variables related to the school's characteristics. First, a dummy variable describes the movement between schools in a given year and takes a value of 1 when a child is observed in two (or more) different schools for two consecutive years. The rest of the outcomes in this second group characterize a school in three dimensions. The first dummy variable, "Public school," takes a value of 1 if a child is enrolled in a public school, and 0 otherwise. A second variable, "Scientific-Humanist," takes a value of 1 if a child who is attending secondary education is enrolled in a school following an academic track to prepare students for college, and 0 otherwise. Finally, since the classmates in the cohort may be observed over the 11 years followed in this study, the rest of the outcomes are useful for measuring the degree of *cream skimming* observed in the school*,* that is, the extent to which schools are able to select better students and remove students at the bottom of the distribution. First, a variable is created with the fraction of students among those coming from other schools who had grades above the median in the previous school. Since a standardized examination comparable across schools is not used here, this variable also helps to assess something about the quality of classmates. Specifically, this variable will increase when the rotation of students into the school is lowered (smaller denominator), or when more of the students who move come from the upper part of the distribution in their previous school (larger numerator). Both of these movements could be related to a higher quality of school. The next dummy variable takes a value of 1 if the GPA in first grade is higher than the median of the students that still remain from the first grade. For a child with a GPA above the median in first grade, this variable will take a value of 1. However, if the school is actively *cream skimming,* this variable will be less likely to take a value of 1 later on. Two dummy variables are also defined that indicate whether or not a particular student has, first, a GPA higher than that of the median of students who have ever moved and, second, a GPA higher than that of the students just moving into the school.[16] Finally, the last outcomes correspond to the average GPA of the rest of the classmates, not counting in this average the student's own GPA.

The descriptive statistics are presented in Table 1. The sample is composed of approximately 250,000 students observed for approximately 10.2 years; 13% of the students attend schools in rural areas and 91% of the children starting primary

[16] These two variables make an attempt to capture whether or not a student is doing better than other students who remain in a given school, or in relation to those just moving into this school. By looking at the evolution of this parameter over time, together with other outcomes, one is able to establish a better picture of the relative return of age at entry and the type of schools these students are moving to over time. Although the absolute impact of this sorting process on average GPA is ambiguous, the relative advantage associated with age at entry should decrease with average quality of classmates. Where some of these schools were actively engaged in *cream skimming*, the probability of having a GPA higher than the median of the students moving into the school should not only be lower but should also fall over time, specifically in relation to the new students in these schools.

Table 1 Descriptive statistics

Attendance	91.00 (15.37)	GPA higher than median grade classmates ever-moved	0.52
Attendance over the median	0.57		
GPA	5.57 (1.06)	GPA higher than median classmates just moving IN to school	0.65
GPA over p50	0.56	Average grade other classmates	5.54 (0.62)
Pass	0.93	Become 6 the year of Entry	0.40
Move school	0.20	Age at entry (days)	2282.63 (115.63)
Public school	0.47	Male	0.51
Scientific or humanistic track	0.16	Same municipality	0.91
Fraction of incoming students over the median in previous	0.52 (0.26)	Class size first grade	32.17
First grade GPA higher than median classmates coming from first grade	0.73	Rural	0.13
		Periods	10.22
		Individuals in first grade	253,794

Standard deviations in parentheses. Standard deviations for proportions are not reported

school do so in a school in their own municipality. The average age at entry is 6.23 years (2282 days), with 40% of the students starting school the year they turn 6. In terms of the selected outcomes, the average attendance in a given year is over 91%. The average GPA is 5.6,[17] with approximately 93% of the students being promoted to the next class in every period. In a given year over the period under analysis, approximately 20% of students change school. In relation to the school type, approximately 47% of the students in a given year are enrolled in a public school. Although approximately 16% of the children are following an academic track, this fraction is driven down for the periods in primary school. When the sample is restricted to students in secondary school, this fraction increases to approximately 60%.[18]

[17] The grade scale in Chile goes from 0 to 7, with 7 being the highest grade. To pass a subject or test, students must obtain a grade of 4 or higher.

[18] The graphic analysis presenting the relationship between each of the outcomes and the student's date of birth is reported in the working paper. For each outcome, a flexible second-degree polynomial is fitted at every side of the four discontinuities. Two observations are noticeable from the figures: first, the existence of a series of discontinuities around the cutoff, and, second, a heterogeneous jump around them.

4 Results

4.1 Discontinuity in the Age at Entry

The RD design as a source of randomization of the treatment not only should ensure that other observed and unobserved factors are uncorrelated with the treatment, but also, and equally importantly, provides a significant variation of the treatment defined as the age at entry to the educational system.[19] Figure 1 presents the source of variation associated with the minimum age at entry using two definitions of the variable of interest: age at entry (in days) and a dummy variable indicating whether or not a child started school in the year she/he turned 6. Firstly, it is observed that those children born at the end of the calendar year are older when starting school and are also less likely to start school the year that they turn 6 (they are more likely to start school the year that they turn 7 or even later). However, conditional on being eligible to start school the year they turn 6 (being born before 1 April or being subject to the same eligibility rule as those born after 1 April but before 1 July), those students born later but to the left of a particular cutoff are the younger ones in their class. Secondly, a distinguishable jump in the age of starting school is observed for children born around 1 April, 1 May, 1 June, and 1 July. However for each of these thresholds a discontinuity in the treatment can be observed; the largest jump, for those children born around the threshold of 1 July, is noteworthy. This large jump around 1 July is explained by the perfect compliance associated with the rule of turning 6 before 1 July. In fact, the fraction of students starting school the year that they turn 6 (as opposed to the alternative of turning 7 or more) drops to practically zero for those children born after 1 July.

Table 2 presents the estimated discontinuities for the "age at entry" (in days) at start of primary school. The results confirm the graphical analysis. Being born after the cutoff date causes entry to be delayed for some students, that is, increases the average age at entry. This impact is significant for each of the cutoffs, but the largest discontinuity is observed for those individuals born after 1 July, who experience an average increase of approximately half a year in the age at entry. The average increase for the rest of the cutoffs is between 15 and 45 days approximately. The results are robust to the selection of the bandwidths, the degree of the polynomial, and the inclusion of other covariates in the specification. Finally, following the

[19] Analysis using an RD design is built on the fact that the variation of the treatment (age at entry) is as good as a randomized assignment for those observations near a particular discontinuity. Formal testing was carried out for discontinuities in baseline characteristics (highest parent's education and gender) for alternative polynomial specifications and different bandwidths. Finally, following McCrary (2008), testing was also carried out for a discontinuity in the distribution of the running variable, by estimating the density of the variable date of birth and formally testing for a discontinuity for each of these cutoffs. First, no evidence was found of a discontinuity in the baseline characteristics. Second, for any of the four thresholds the null hypothesis which supports the graphical analysis can be rejected, and there is no evidence of a precise manipulation of the date of birth. For reasons of space, these checks were left in the working paper version.

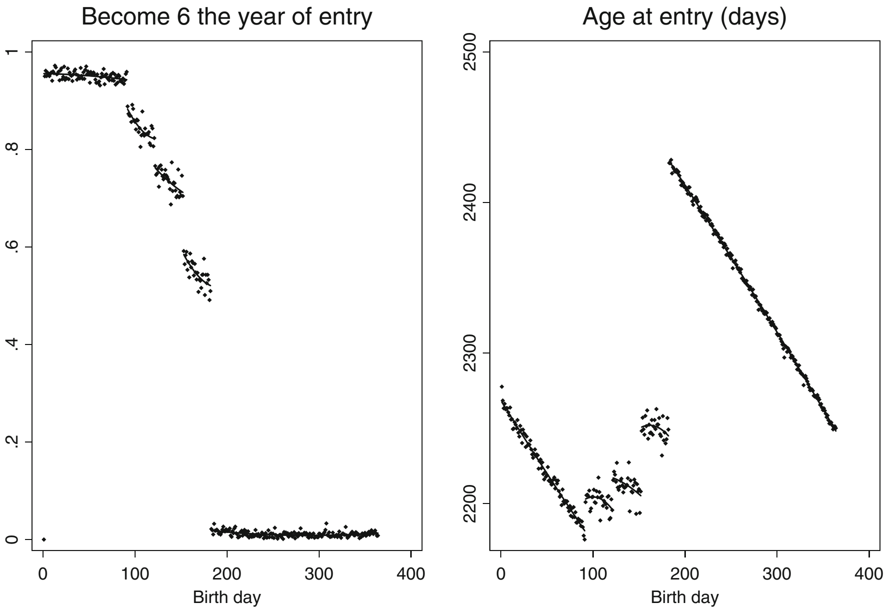


Fig. 1 Age at entry and minimum age entry rule

equivalence with an IV approach, the value of the F-statistic for the null of the relevance of the excluded instruments is large enough to disregard any concern about weak instruments in all the specifications and selected bandwidths.

4.2 Impact of Age at Entry on Selected Outcomes

The impact of age at entry on the two selected groups of outcomes is reported in Table 3. The overall picture from Table 3 regarding the impact of increasing age at entry on the different school achievement outcomes is not only that it is positive but also that the impact is still present 11 years after starting primary school. The only exception regarding this positive impact of increasing age at entry is observed for some of the outcomes (attendance) for the ninth year after school entry, which would correspond to the first year of secondary school for those children who had not repeated a grade.

Specifically, first, for the variable "attendance" it is observed that children with a higher age at entry increase their attendance by between 1 and 2 percentage points during the first 2 years after starting primary school and the second year of secondary school. This impact in terms of school days, given an average attendance of 91% in the population, means that increasing age at entry by 1 year is associated with approximately 2–5 more days of classes during a specific school year. The

Table 2 Impact of minimum age entry rules on age of entry (days)

		1			2			3			4		
		15 days	10 days	7 days	15 days	10 days	7 days	15 days	10 days	7 days	15 days	10 days	7 days
I unconditional	Born after April 1st	28.3233***	27.5121***	31.8843***	28.7104***	37.9610***	40.2898***	29.4427***	38.8296***	41.7091***	39.4482***	40.0230***	40.9451***
		[3.4283]	[4.4860]	[4.1711]	[5.1742]	[3.9449]	[5.0254]	[5.7237]	[4.6506]	[5.9641]	[6.3781]	[5.7847]	[6.3855]
	Born after May 1st	21.3254***	22.1842***	22.6974***	25.9690***	26.3100***	25.9260***	28.4573***	28.4133***	29.3772***	26.0146***	25.2956***	26.9640***
		[3.6022]	[3.8046]	[4.3265]	[5.2597]	[5.4136]	[6.1072]	[6.2929]	[6.3891]	[6.8575]	[7.2710]	[6.9952]	[7.1053]
	Born after June 1st	47.7396***	47.4474***	49.4726***	47.8233***	51.3239***	47.2595***	50.6623***	50.8338***	48.0819***	54.8340***	52.2654***	61.4692***
		[4.3010]	[4.6499]	[5.5082]	[6.9130]	[7.6219]	[8.4934]	[10.1476]	[10.8000]	[12.9671]	[13.3338]	[14.3005]	[16.8077]
	Born after July 1st	174.1180***	175.7277***	174.5104***	175.2875***	173.3698***	176.7629***	176.7504***	169.1098***	172.3772***	169.5726***	171.5069***	163.7984***
		[2.8339]	[3.0380]	[3.1186]	[4.0611]	[3.8915]	[4.6390]	[5.5525]	[4.8457]	[5.8295]	[6.5500]	[5.9557]	[7.7617]
	Observations	62,937	43,288	30,791	62,937	43,288	30,791	62,937	43,288	30,791	62,937	43,288	30,791
	R-squared	0.2001	0.2097	0.2107	0.2002	0.2099	0.2107	0.2002	0.2099	0.2108	0.2003	0.2099	0.2109
	F-statistic	1086.93	939.09	858.82	625.6	765.79	488.96	321.76	390.92	231.66	242.18	284.4	179.18

II conditional	Born after April 1st	29.5483***	28.7123***	33.0564***	29.5923***	39.5316***	39.6229***	29.4691***	39.5520***	40.3621***	40.1387***	39.6881***	38.3298***
		[3.3503]	[4.4369]	[3.9279]	[5.2206]	[3.9459]	[5.1033]	[5.8287]	[4.7182]	[6.0876]	[6.1773]	[5.6223]	[6.0367]
	Born after May 1st	22.1573***	22.8459***	23.0487***	26.6489***	27.3817***	27.3359***	27.8893***	27.9771***	29.7901***	27.7967***	27.9226***	31.2337***
		[3.6029]	[3.8502]	[4.3750]	[5.2151]	[5.3175]	[5.9482]	[6.0016]	[6.1382]	[6.6256]	[7.2896]	[7.0224]	[7.1966]
	Born after June 1st	47.7814***	48.0355***	49.6617***	48.2743***	50.8721***	46.6316***	51.0012***	50.0685***	47.3477***	52.7347***	50.0947***	59.1699***
		[4.1109]	[4.4200]	[5.2006]	[6.5390]	[7.2403]	[8.0494]	[9.7263]	[10.2374]	[12.3223]	[12.5745]	[13.3697]	[15.9308]
	Born after July 1st	173.1264***	174.6430***	173.2580***	174.0252***	172.6457***	176.4671***	175.5515***	169.0771***	172.1695***	168.9480***	170.6162***	161.7350***
		[2.8338]	[3.0815]	[3.1847]	[4.0713]	[3.7673]	[4.4782]	[5.1979]	[3.9867]	[5.4183]	[5.4995]	[5.4193]	[6.7335]
	Observations	62,922	43,281	30,786	62,922	43,281	30,786	62,922	43,281	30,786	62,922	43,281	30,786
	R-squared	0.2254	0.2343	0.2367	0.2254	0.2345	0.2367	0.2255	0.2345	0.2367	0.2255	0.2345	0.2368
	F-statistic (excluded instruments)	1095.64	938.02	855.54	638.36	769.77	511.2	327.08	424.72	236.28	274.84	306.31	191.38

Robust standard errors in brackets ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$. Standard errors are clustered at birthday level

Other covariates in the conditional specification are dummies for male, rural status and whether or not the child started school in the same municipality where she/he resided

Table 3 Impact of age entry on selected outcomes

Years since first enrollment	Attendance	Attendance over median	GPA	GPA over median	Pass	Move school	Public school	Scientific-humanistic track	Fraction of incoming students over the median in previous school	First grade GPA higher than median classmates coming from first grade	GPA higher than median grade classmates evermoved	GPA higher than median classmates just moving IN to school	Average grade other classmates
1	1.9691 * *	0.1077***	0.4793***	0.3106***	0.0163		−0.0725***			0.3043***	0.1448***		0.003
	[0.8130]	[0.0231]	[0.0524]	[0.0243]	[0.0104]		[0.0256]			[0.0262]	[0.0142]		[0.0277]
2	1.1253	0.0117	0.3966***	0.2037***	0.0546***	−0.0271	−0.0755 * *		−0.0119	0.2246***	0.1641***	0.1593***	0.0736***
	[0.7038]	[0.0250]	[0.0500]	[0.0241]	[0.0106]	[0.0172]	[0.0297]		[0.0284]	[0.0296]	[0.0253]	[0.0157]	[0.0186]
3	1.1880 * *	0.0321	0.3117***	0.1770***	0.0486***	0.0071	−0.0286		0.006	0.1370***	0.1752***	0.1571***	0.0151
	[0.5734]	[0.0277]	[0.0415]	[0.0261]	[0.0099]	[0.0227]	[0.0352]		[0.0143]	[0.0279]	[0.0242]	[0.0354]	[0.0199]
4	0.1595	−0.0259	0.2160***	0.1467***	0.0206 * *	−0.0307*	−0.0814 * *		0.0066	0.1611***	0.1242***	0.0823***	0.0062
	[0.5310]	[0.0238]	[0.0403]	[0.0226]	[0.0090]	[0.0184]	[0.0349]		[0.0140]	[0.0265]	[0.0298]	[0.0240]	[0.0211]
5	0.5353	0.0374	0.2060***	0.1355***	0.018	0.0470***	−0.0830 * *		0.0021	0.1262***	0.1184***	0.1240***	−0.0015
	[0.4952]	[0.0275]	[0.0402]	[0.0262]	[0.0117]	[0.0173]	[0.0392]		[0.0148]	[0.0204]	[0.0297]	[0.0348]	[0.0212]
6	0.1043	0.0443	0.2199***	0.1503***	0.0259 * *	−0.0199	−0.0722 * *		0.0065	0.1837***	0.1169***	0.0989***	0.0015
	[0.6490]	[0.0282]	[0.0432]	[0.0184]	[0.0106]	[0.0174]	[0.0337]		[0.0142]	[0.0201]	[0.0189]	[0.0219]	[0.0223]
7	0.2846	0.0853***	0.1696***	0.0880***	0.0263 * *	0.0031	−0.0727 * *		0.0138	0.1611***	0.0848***	0.0650***	0.0639***
	[0.5740]	[0.0273]	[0.0359]	[0.0275]	[0.0131]	[0.0246]	[0.0327]		[0.0114]	[0.0193]	[0.0209]	[0.0193]	[0.0199]
8	0.4302	0.0244	0.0774 * *	0.0945***	−0.0021	−0.0287	−0.0823***		0.0006	0.1461***	0.0790***	0.0643 * *	−0.0231
	[0.6122]	[0.0359]	[0.0328]	[0.0221]	[0.0109]	[0.0201]	[0.0312]		[0.0121]	[0.0319]	[0.0241]	[0.0268]	[0.0232]
9	−3.3135***	0.0662***	0.0597	0.0661***	−0.0225	−0.0213	−0.0648***	0.1457***	0.0241*	0.1056***	0.0488 * *	0.0337	−0.1122 * *
	[0.9667]	[0.0242]	[0.0520]	[0.0229]	[0.0154]	[0.0241]	[0.0215]	[0.0249]	[0.0126]	[0.0172]	[0.0236]	[0.0304]	[0.0449]
10	2.1522*	−0.0312	0.1335 * *	0.0643***	0.0226	0.0627***	−0.0629 * *	0.1365***	0.0347 * *	0.1116***	0.0699 * *	0.0769***	0.0484
	[1.2364]	[0.0245]	[0.0617]	[0.0214]	[0.0188]	[0.0199]	[0.0246]	[0.0263]	[0.0158]	[0.0158]	[0.0273]	[0.0281]	[0.0462]
11	−0.6063	0.0392	0.1504*	0.0724***	0.0259	−0.1395***	−0.0548 * *	0.0947***	0.0449***	0.0716***	0.0751***	−0.0056	0.1129 * *
	[1.6025]	[0.0297]	[0.0836]	[0.0197]	[0.0174]	[0.0232]	[0.0242]	[0.0256]	[0.0115]	[0.0169]	[0.0215]	[0.0212]	[0.0472]

Robust standard errors in brackets $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Standard errors are clustered at birthday level

Other covariates in the specification are dummies for male, rural status and whether or not the child started school in the same municipality where she/he resided exception in the size of the impact of age at entry is found for the first year of secondary school, when attendance falls by an average of 3 percentage points. However, even in the ninth year after starting school, by looking at the impact on the probability of having an attendance rate higher than the median in the class, one can discern a positive effect associated with age at entry. This finding suggests that the negative effect on attendance rate captured during the first year of secondary school reflects some school effects rather than an individual reduction in attendance.

Second, a higher age at entry is associated with a higher annual GPA for all years for which this cohort of students is followed. However, this impact tends to diminish with time. Specifically, the impact on average GPA falls from approximately 0.5 points for a child during her/his first year in the school system to 0.15 points 11 years after entry into the system (which would be the third year of secondary school for those children who have not repeated a grade). These magnitudes are not only statistically significant but also economically important. The magnitude of the increase in GPA (0.5 or 0.15 points) is equivalent to the difference between a student on the median of the GPA distribution and one on the 75th (60th) percentile of this distribution. Can this observed impact on GPA be driven by differences in grade inflation between schools, or differences in the requirements established in schools? The results for the outcomes indicating the place on the distribution of GPA in relation to the students within the same school cohort do not support this hypothesis. A higher age at entry increases by almost 31 percentage points the likelihood that a child's GPA is above the median in their first year of school, and increases it by approximately 7 percentage points 11 years after school entry.[20] Consistent with this positive impact on GPA, it is found that a higher age at entry is associated with an increase in the probability of passing a grade of between 2 and 5 percentage points.

In this way, and in contrast to other studies, these results show an impact of age at entry that is still observed 11 years after the start of primary school. Tincani (2014) shows that, in Chile, private schools (privately and voucher funded) not only specialize in higher-ability students, but are also able to attract higher-quality teachers from public schools and from other sectors, by offering them better salaries. In relation to the current problem, it has been shown that a higher age at entry provides some advantage in terms of school achievement. Thus, in light of Ticani's findings, these better students will be more likely (at some point in their school life) to sort into better schools. Although the absolute impact of this sorting process on average GPA is ambiguous,[21] the *relative* advantage associated with delayed entry

[20] Although the same qualitative results are observed for the other variables indicating whether or not a student is above the 75th, 50th, or 25th percentiles, the strongest impacts are observed in the upper part of the distribution of the GPA. In fact, only in the first years of school life it is observed that a higher age at entry increases the probability of being above the 10th percentile in the GPA distribution of the class.

[21] Anecdotal evidence suggests higher grade inflation among worse schools (http://www.uchile.cl/portal/presentacion/historia/luis-riveros-cornejo/columnas/5416/inflacion-de-notas), that is, the effect of age at entry should be negatively correlated with grade inflation, and the impact of age at entry will be more likely to be observed among better schools. If, on the other hand, better schools

should decrease with average quality of classmates. The rest of the analysis seeks to measure the effect of age at entry on schools' characteristics and movement between schools.

Mixed results are observed for the outcomes characterizing movement between schools. Although children whose entry to school is delayed are less likely to change schools in the fourth year after the start of primary school, a year later they are more likely to move to other schools. In fourth grade, children take a National Examination (SIMCE). This examination aims to measure the quality of the education provided by schools, and the results might later be used by parents to choose schools. Moreover, there is anecdotal evidence indicating that schools have some leeway in selecting which students take this examination. Along these lines, this finding that a higher age at entry is associated not only with better grades but also with being less likely to change school in the year when this national examination takes place suggests that schools might have some power to retain these better students, at least in the short run (the year that this examination takes place). Also, during the first year of secondary school it is observed that delayed entry is associated with an increase in the chance of changing schools. In fact, an additional year in child's age at school entry is associated with an increase of approximately 6 percentage points in the likelihood of changing school between the ninth (first year of secondary school) and tenth years after starting primary school. The start of secondary school in Chile is characterized by approximately 50% of the students in the educational system switching schools. This amount of friction might induce some students to actively search for a new school. Where this search effort is positively correlated with early educational outcomes, it would explain why higher age at entry results in an increased search effort in the periods with higher friction in the system. In fact, this greater search effort is consistent with a reduction in the probability of switching schools between the tenth and eleventh years. Also consistent with this increase in the fraction of students switching school at the beginning of secondary school is the already reported drop in school attendance during the first year of secondary school.

Regarding school type, it is first observed that a higher age at entry is associated with a decrease in the probability that a child attends a public school for almost all years under analysis. Second, it is observed that delaying school entry increases the likelihood that a child will follow an academic track by approximately 13 percentage points. In terms of the sample, this means that this last impact corresponds to an approximately 25% increase in the fraction of students in the educational track that aims to prepare students for college.

For the outcomes characterizing the school measured by the classmates choosing the school (who have moved there), it is observed that increasing the age at entry is associated with a rise in the fraction of students arriving from the upper part of the GPA distribution in their previous school, but specifically in secondary school.

set higher standards, it will be harder to observe a stronger effect of age at entry as children move to better schools.

Therefore, the results reveal not only that higher age at entry is associated with a lower probability of being enrolled in schools linked to lower-quality teachers, as observed in public schools (Tincani 2014), but also that these students have better-quality classmates in secondary school, where they are more likely to follow an academic track. In fact, it can be assumed that all of these factors contribute to the drop in the GPA in secondary school, due to a tougher academic track and the lower likelihood of grade inflation reported in better schools. It is worth noting that the timing of these impacts is consistent with the drop in the GPA observed when starting secondary school. The last four outcomes explore the impact of age at entry on the probability of being enrolled in a school that is actively *cream skimming*. It has been shown that children with a higher age at entry have a greater likelihood of having a GPA higher than the mean of their classmates. Where some of these schools were actively engaged in *cream skimming*, however, the probability of having a GPA higher than the median of the students moving into the school should not only be lower but also fall over time. This is what is observed from the impact of age at entry on the outcomes that measure the probability of having a GPA higher than the median of students moving into the school in a given year, or of those who have moved to the school at some point in the past. That is, it is observed that age at entry increases the probability of having a GPA higher than the median of students moving into the school, but this increase in the probability is lower than the increase in the probability of having a GPA higher than the median of all students in the class (fourth column in Table 3) for almost all the years. Secondly, over the school life this probability decreases and by the time secondary school is reached it is not statistically significant. Also consistent with the hypothesis that age at entry increases the likelihood of being enrolled in a school actively *cream skimming*, analysis of the GPA in first grade over the years, but with the mean of the classmates arriving from first year of primary school, shows that this probability decreases over time. Finally, it is observed for some years, and with the exception of the first year of secondary school, that age at entry increases the probability of being enrolled in a school where classmates have, on average, a higher GPA.

5 Conclusions

The findings of this study confirm not only that delaying school entry has a positive effect on GPA, attendance and the likelihood of passing a grade but also that this impact tends to wear off across time. Nevertheless, in contrast to previous studies, the findings reveal that this impact can still be observed 11 years after a child has started school. Moreover, evidence of the effect of age at entry on school type provides a potential explanation for the fading of the impact on academic achievement throughout the school life. Specifically, a higher age at entry decreases the likelihood that a child will be enrolled in municipal (public) schools, which are characterized by a less active selection of students and lower-quality teachers. Consistent with these differences in academic selection (competition), children with a higher age at entry have a higher probability of being enrolled in schools where the children arriving from other schools were in the upper GPA distribution in their previous school. Evidence is also found that age at entry has a positive effect on the likelihood that a child follows an academic track in high school. Finally, evidence is found that age at entry is associated with an increase in the probability that a child will be enrolled in a school actively engaged in *cream skimming*, which also explains the drop in the impact of age at entry on school achievements.

Acknowledgements We want to thank seminar participants at the Departments of Economics at Carlos III and Universidad Alberto Hurtado for comments and suggestions. Julio Cáceres-Delpiano gratefully acknowledges financial support from the Spanish Ministry of Education (Grant ECO2009-11165), Spanish Ministry of Economy and Competitiveness (MDM 2014-0431), and Comunidad de Madrid MadEco-CM (S2015/HUM-3444).

References

Angrist JD, Kruger AB (1991) Does compulsory schooling attendance affect schooling and earnings? Q J Econ 106:979–1014

Berthelon M, Kruger D (2012) Risky behavior among youth: Incapacitation effects of school on adolescent motherhood and crime in Chile. J Public Econ 95(1–2):41–53

Black S, Devereaux P, Salvanes P (2011) Too young to leave the nest? The effect of school starting age. Rev Econ Stat 93(2):455–467

Buckles K, Hungerman D (2013) Season of birth and later outcomes: old questions, new answers. Rev Econ Stat 95(3):711–724

Cascio EU, Lewis EG (2006) Schooling and the armed forces qualifying test. Evidence from school-entry laws. J Hum Resour 41(2):294–318

Contreras D, Rau T (2012) Tournament incentives for teachers: evidence from a scaled-up intervention in Chile. Econ Dev Cult Chang 61(1):219–246

Contreras D, Larranaga O, Flores L, Lobato F, Macias V (2005) Politicas educacionales en Chile: vouchers, concentracion, incentivos y rendimiento. In: Cueto S (ed) Usos e impacto de la informacion educativa en America Latina. Preal, Santiago, pp 61–110

Deming D, Dynarski, S (2008) The lengthening of childhood. J Econ Perspect 22(3):71–92

Gauri V, Vawda A (2003) Vouchers for basic education in developing countries. A principal-agent perspective. World bank policy research working paper 3005. World Bank, Washington

Hahn J, Todd P, van der Klaauw W (2001) Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69(1):201–209

Hsieh C, Urquiola M (2006) The effects of generalized school choice on achievement and stratification: evidence from Chile's voucher program. J Public Econ 90(8–9):1477–1503

Kelly B, Dhuey E (2006) The persistence of early maturity: International evidence of long-run age effects. Q J Econ 121(4):1437–1472

Lee DS, Lemieux T (2010) Regression discontinuity designs in economics. J Econ Lit 48:281–355

McCrary J (2008) Manipulation of the running variable in the regression discontinuity design: a density test. J Econ 142(2):698–714

McEwan P, Shapiro J (2006) The benefit of delayed primary school enrollment. Discontinuity estimates using exact birth dates. J Hum Resour 43(1):1–29

Mizala A, Urquiola M (2013) School markets: the impact of information approximating schools? effectiveness. J Dev Econ 1003:313–335

Sandler M (2011) The effects of maternal employment on the health of school-age children. J Health Econ 30(2):240–257

Stipek D (2002) At what age should children enter kindergarten? A question for policy makers and parents. Soc Policy Rep 16(2):3–16

Tincani MM (2014) School vouchers and the joint sorting of students and teachers. HCEO Working Paper No. 2014-02. Human capital and economic opportunity global working group, University of Chicago

Julio Cáceres-Delpiano after earning his PhD in Economics from University of Maryland (USA) in 2005, joined the Department of Economics at Universidad Carlos III de Madrid (Spain), where he currently holds a position as an Associate Professor (with tenure). He has published in the areas of labor, population, and family economics. Publications in international journals include outlets such as the Journal of Human Resources, Demography, Journal of Labor Economics, Journal of Health Economics, and The B.E. Journal of Economic Analysis & Policy.

Eugenio P. Giolito holds a PhD in Economics from University of Maryland (USA) and is currently Associate Professor at Universidad Alberto Hurtado (Chile). Previously, he was Assistant Professor at Universidad Carlos III de Madrid (Spain). He has published in the areas of labor, population, and family economics. Publications in international journals include outlets such as the International Economic Review, Journal of Labor Economics, and The B.E. Journal of Macroeconomics.

Part IV
Use of Results

Use of Administrative Data for Counterfactual Impact Evaluation of Active Labour Market Policies in Europe: Country and Time Comparisons

Beatrice d'Hombres and Giulia Santangelo

1 Introduction

Fostering access to administrative data is without doubt one prerequisite to make policy evaluation more systematic. Administrative data have the advantage of providing detailed and accurate information on large samples of individuals, over repeated periods of time. Since administrative information is often collected for management and monitoring purposes, the possibility of using this data source for promoting evidence-based policy-making might also be more cost efficient than the use of alternative sources of information such as survey data.

This chapter documents how widespread is the use of administrative data for counterfactual impact evaluation (CIE) of active labour market policies (ALMPs). The analysis is based on articles and working papers published since 2000 that evaluate the impact of ALMPs implemented in Europe. After documenting differences across countries and over time in the use of CIE-based evidence of ALMPs, this chapter discusses how data availability might affect the choice of the econometric methods employed for measuring the causal effect of the underlying intervention. It is argued that the data source, whether administrative or survey based, correlates with the comprehensiveness of the CIE. In other words, the feasibility of measuring the impact of the intervention on an array of outcomes, comparing the effectiveness of alternative interventions and making comparisons across different treated groups, seems to be related to the type of data used.

Electronic supplementary material. The online version of this chapter (https://doi.org/10.1007/978-3-319-78461-8_17) contains supplementary material, which is available to authorized users.

B. d'Hombres · G. Santangelo (✉)
European Commission, Joint Research Centre (JRC), Ispra, Italy
e-mail: beatrice.dhombres@ec.europa.eu; giulia.santangelo@ec.europa.eu

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_17

While various meta-analyses have been carried out to assess the effectiveness of ALMPs (Card et al. 2010; Kluve 2010; Bratu et al. 2014; Card et al. 2017), this is, to the best of the authors' knowledge, the first study to document a link between administrative data availability and CIEs' use and comprehensiveness.

The analysis presented in this chapter is largely drawn from the report "Knowledge gaps in evaluating labour market and social inclusion policies" (Bratu et al. 2014) and the associated online Counterfactual Evaluation Archive (CEA).[1] The Centre for Research on Impact Evaluation (CRIE) of the Joint Research Centre of the European Commission was commissioned to write this report by the Directorate-General for Employment, Social Affairs and Inclusion. It reviews evidence on the impact of labour market policies of the type funded by the European Social Fund. The focus on ALMPs was a response to the need for deeper knowledge about the results of the interventions implemented in Europe to address a wide range of labour market problems, such as youth unemployment and social inclusion. In the light of both the increasing focus of the European Commission on policy effectiveness and of the tightened national budgets for the current 2014–2020 programming period, the ultimate objective of the report was therefore to identify possible areas of priority for the CIE of the interventions funded through the European Social Fund.

On the basis of the results of this report, in 2015, the CRIE launched the CEA, to summarise information on what works for whom in the areas of employment and social inclusion. The CEA, an online database which is regularly updated, compiles information on published articles and working papers related to CIE of ALMPs. The discussion in this chapter is based on the most recent update of the CEA. More precisely, the information is drawn from 111 CIEs of ALMP intervention-based papers published over the period January 2000–October 2016. The interventions evaluated took place in the EU-28.

The remainder of the chapter is structured as follows. Section 2 explains the data collection protocol, while Sect. 3 discusses the main findings and lessons learnt from the analysis of the CEA, and Sect. 4 concludes.

2 Data Collection

2.1 Definition of the Active Labour Market Policies Included in the Analysis

Three broad categories of interventions, corresponding to the classification of ALMPs used by both Eurostat and the Organisation for Economic Co-operation and Development, have been considered in the analysis below. These are (1) training, (2) employment incentives and (3) labour market services.

[1] https://crie.jrc.ec.europa.eu/CIE_database/cieDatabase.php.

Trainings encompass measures aimed at enhancing the human capital of participants. A distinction is made between (a) classroom/vocational training, (b) on-the-job training and (c) other types of training that are neither classroom/vocational training nor on-the-job training (e.g. training for auxiliary/basic skills).

Employment incentives aim at offering work experience to unemployed individuals to keep them in contact with the labour market. Employment incentives support both the private and the public sectors. Private sector employment incentives include hiring and wage subsidies. Start-up incentives (e.g. tax incentives) or grants, which help unemployed individuals to start their own business, also fall under this category. Public sector employment incentives include public employment schemes to create jobs in public firms or in public activities that produce public goods or services.

Labour market services aim at supporting the unemployed throughout their job search process. They include a wide range of forms of assistance for job seekers, such as career management, job boards, job search assistance, career coaching and job referral and placement. For the CEA, in line with the literature, labour market service interventions have been divided into the following sub-interventions: (a) job search assistance, (b) counselling and monitoring and (c) job placement and relocation assistance.

2.2 Search Strategy: Identification of the Literature Examining the Effect of Active Labour Market Policies

Four online databases were searched for relevant academic articles: (1) Scopus, (2) RePEc IDEAS, (3) SSRN and (4) IZA Discussion Papers Database.[2] The protocol used to identify the relevant studies involved the following three steps.

First, within each database a search was carried out using the terms ("labour market" OR "labour market" OR "job") AND ("evaluation" OR "impact" OR "data" OR "intervention" OR "program"), to broadly identify studies related to labour market and evaluation. Each set is connected by AND, while individual search terms within a set are connected by OR.

Intervention-specific keywords[3] and the publication period (between January 2000 and October 2016) were also added to this search process. At this stage there was no restriction in the query regarding the target groups or the evaluation methods, to ensure that the search would deliver the most comprehensive set of results along these dimensions.

Second, given the large number of papers that the search identified, the results were further filtered by title. Third, the studies in the resulting list were scrutinised

[2] See http://www.scopus.com, http://ideas.repec.org, http://www.ssrn.com and http://legacy.iza.org/en/webcontent/publications/papers for additional information.

[3] See Bratu et al. (2014) for additional information on the keywords used for the search.

to select those to be included in the CEA.[4] At this stage, a thorough analysis of potentially relevant studies was carried out to identify the intervention types, outcome variables, evaluation methods and target groups.

In order to be selected, the evaluation had to (1) be based on CIE methods (regression, randomisation, propensity score matching, difference in differences, regression discontinuity design and instrumental variable methods), (2) be focused on interventions aiming at individuals[5] and (3) examine the impact of ALMPs on specific labour market outcomes, namely, employment status, duration of employment or income/wage level. The studies included in the archive are all reported in the online appendix. For a brief explanation of CIE, see Box 1 below.

Box 1: Counterfactual Impact Evaluation in a Nutshell

The purpose of a CIE is to assess the causal relationship between participation in an intervention and the outcome of interest (e.g. employment probability). It therefore requires comparing the participants' outcome in the actual situation with that which would have occurred had they not participated. This is called the *fundamental evaluation problem*, as it is impossible to observe the outcomes of the participants in the latter situation, i.e. in the counterfactual state. Therefore, it is necessary to find an adequate *control group* of non-participants with which the participant group can be compared.

The two groups should be as similar as possible so as to ensure that the difference in their outcomes can be causally attributed to the intervention. Usually, groups of participants and non-participants are different in dimensions other than participation, either because the intervention is targeted at a particular group of individuals or because individuals self-select into the intervention. In either case, this leads to selection bias, making it misleading to simply compare the outcomes of participants and non-participants.

CIE methods are statistical and econometric techniques to compare the outcomes of participants and non-participants, while taking into consideration the selection bias problem by controlling for pre-existing differences between the two groups. The most compelling way to tackle the selection bias is the experimental setting (randomised control trial (RCT)), whereby individuals are allocated randomly to one of the two groups (participants or non-participants). This eliminates the selection bias problem because, given the randomised assignment, the two groups are similar in all respects but the intervention participation. RCTs however have limited external validity, i.e.

(continued)

[4] Note that papers included in the CEA are evaluations of interventions aimed at individuals, even if the interventions involved or targeted firms. Examples of such interventions are training or hiring subsidies.

[5] In case the intervention targets firms, the units treated should be the employees, for instance, through training or hiring subsidies.

results cannot easily be generalised to different contexts. When experiments are not economically or ethically feasible, nonexperimental evaluation methods can be applied to ensure the comparability of groups. Nonexperimental methods include *regression*[6] and *matching* methods, *difference in differences*, *regression discontinuity design* and *instrumental variables*. These approaches consist of statistical methods to control for the selection bias so as to be able to identify a proper comparison group.

2.3 Article Coding: Collected Information

For each selected paper, the following features were coded: the country where the intervention took place, the year of the intervention, the target population (unemployed, young unemployed, disadvantaged young unemployed, elderly unemployed, long-term unemployed, low-skilled unemployed, employed, inactive, disabled and women), the evaluation method (propensity score matching, regression discontinuity design, instrumental variables, difference in differences, regression) and the data used (administrative and/or survey) for evaluating the impact of the underlying intervention.[7] For the purpose of this chapter, additional information was also coded, relating to the data quality (sample size and the number of data sources used in the empirical analysis) and the comprehensiveness of the CIE analysis (number of outcome indicators considered, a measure of the intervention impact on different groups of participants).

3 Findings

3.1 Active Labour Market Policies Subject to Counterfactual Impact Evaluation Studies, Target Groups and Outcome Indicators

Applying the search protocol resulted in the identification of 111 relevant papers, among which, as shown in Fig. 1, 30.6% evaluate training interventions, 30.6%

[6] Regression methods can be considered similar to matching methods as they rely on the same identification assumption, i.e. all differences between participants and non-participants are observable to the evaluator (selection on observables or conditional independence assumption). Accordingly, this method is included in the CIE category. However, propensity score matching is generally a more appropriate CIE method because, in contrast to regression, it is based on non-parametric estimation procedure and includes a check for the common support assumption.

[7] Additional collected information includes the funding source of the intervention being evaluated, the authors, the digital object identifier (doi) and the publication year.

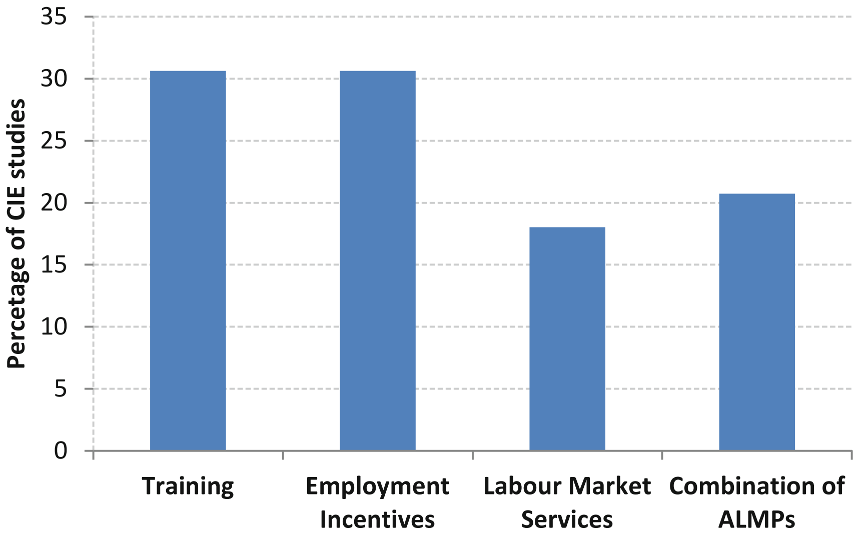


Fig. 1 Counterfactual impact evaluation studies and categories of active labour market policies. Source: https://crie.jrc.ec.europa.eu/CIE_database/cieDatabase.php; authors' calculations

examine the impact of private and public employment incentives and 18% study the effect of labour market services interventions. The remaining 20.7% measure the effect of more than one ALMP.

Around 12% of CIE studies examining training interventions look at on-the-job training programmes, while the other interventions relate to classroom/vocational training (52.9%) or to a combination of these two categories of intervention (35.3%). The majority of CIEs on employment incentives assess the impact of private sector incentives (64.7%), while only 14.7% specifically evaluate incentives in the public sector and 20.6% assess both sets of interventions. In 50% of the cases, CIEs of labour market services are related to job search assistance programmes. Counselling and monitoring make up 20% of these CIEs, whereas the remainder falling under this category includes an assessment of several labour market services at the same time.

Most of the interventions target unemployed individuals (91%). In addition, among these studies, those targeting subcategories of unemployed individuals often concentrate on the long-term unemployed and the young unemployed.[8] Only 7.2% of the interventions focus on employed individuals. The most common outcome indicators used in the CIEs measure the labour market status of the participants and non-participants, namely, the employment rate at the end of the intervention or the

[8] The evidence on the effectiveness of ALMPs targeting other disadvantaged groups such as the elderly unemployed, the low-skilled unemployed or the inactive is scant because of the small number of interventions that address these groups.

rate of exit from unemployment at different periods of time following the conclusion of the intervention.

Around 61.3% of the CIEs selected for the CEA examine the impact of an intervention on more than one outcome indicator. Occasionally this depends on the availability of longitudinal data, i.e. on whether or not information on the labour market history of the population targeted by the intervention is accessible. Such longitudinal information allows an estimation of the intervention effect on employment status at repeated periods in time. In other instances, CIE studies examine the effect of the intervention on both employment and income outcome indicators.

Almost 69.4% of the studies test the presence of heterogeneous effects across population groups (by age, such as young and elderly, or by gender). More specifically, the impact of the underlying intervention is frequently measured separately for males and females and/or for different regional locations within a given country.

3.2 Distribution of Counterfactual Impact Evaluation Studies Across EU Countries and Authorship

The 111 studies, published between January 2000 and October 2016, estimate the effect of interventions taking place in 19 different EU-28 countries, with Germany being by far the country where most of the interventions subject to a CIE took place (51 of the 111 CIEs). The preponderance of CIE studies in Germany can be linked with the introduction of the so-called Hartz labour market reforms, which took place between 2003 and 2005, the evaluation of which the German government commissioned from a number of research institutes (Jacobi and Kluve 2006). As shown in Fig. 2, Sweden and Denmark rank second and third, with ten and eight CIEs, respectively. In general, there is a contrast between the number of CIE studies in West and East European countries. Six, five and four CIE studies were found for France, Italy and the United Kingdom, respectively. CIEs have also been carried out in Poland, Romania, Slovakia, Slovenia, Bulgaria and Latvia, but generally only one CIE study could be found per country.[9] No interventions subject to a CIE were found in Greece, Estonia or Lithuania. The majority of German studies concern training interventions and employment incentives. CIEs of labour market services are more evenly distributed across the EU countries.

The analysis of scientific collaborations in Germany helps to provide a better understanding of the distribution of CIE studies within the country.

For this, as in Newman (2001, 2004a, 2004b) and Ball et al. (2013), network analysis is employed to examine the structure of collaboration networks on the basis of individuals' co-authorship. In Fig. 3 two authors are considered connected

[9] Two studies have been identified in the case of Poland.

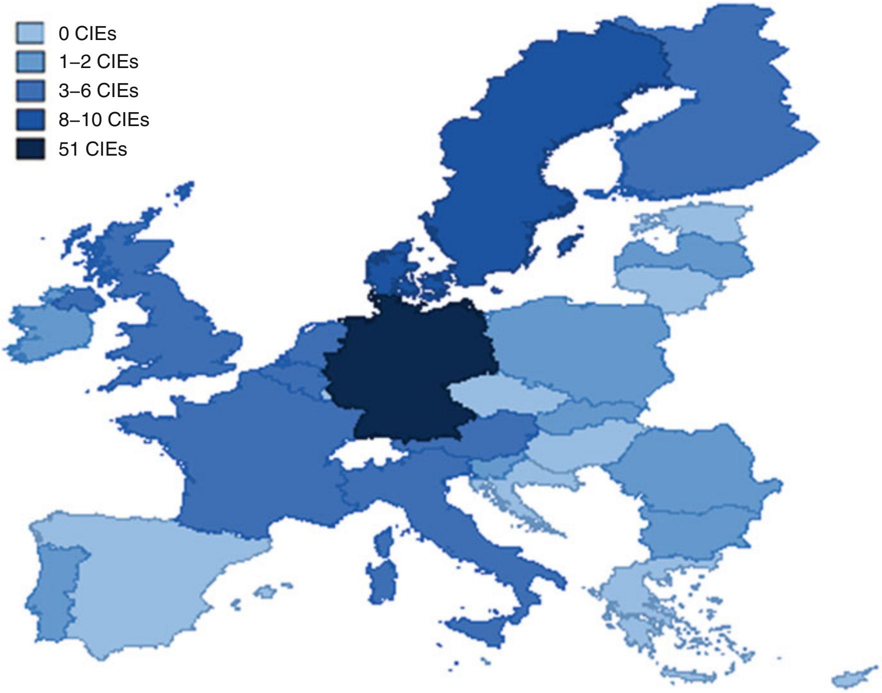


Fig. 2 Number of counterfactual impact evaluation studies per country. Source: https://crie.jrc.ec.europa.eu/CIE_database/cieDatabase.php; authors' calculations

if they have co-authored at least one paper. The strength of collaborative ties is measured on the base of the number of papers co-authored by pairs of authors and is represented by the thickness of the lines connecting them. The size of the nodes is proportional to the number of papers authored by each researcher. Although in a merely visual way, Fig. 3 provides a proxy of how concentrated the authorship of CIE studies is. Names are shown for the authors who have each published more than two CIE studies (maximum 13) based on German data. The concentration of studies among a few core authors may depend on several factors, such as an established tradition in specific departments of performing CIE of ALMPs or the opportunity to access more widely administrative data, in particular for the evaluation of the Hartz reforms. Since most studies build on the IZA/IAB Administrative Evaluation Dataset provided by the Institute for Employment Research (See Eberle and Schmucker 2015), the connectedness within the co-authorship networks and the affiliation of the researchers (at the time of the CIE studies) clearly underline the importance of access to these data for any CIE of ALMPs in Germany. Since the IAB database has been opened to researchers from outside Germany, CIEs of ALMPs in Germany tend to be less concentrated among only a few authors. This is a first indication that the availability of high-quality administrative data is probably related to the application of counterfactual methods for data-driven evidence-based policy.

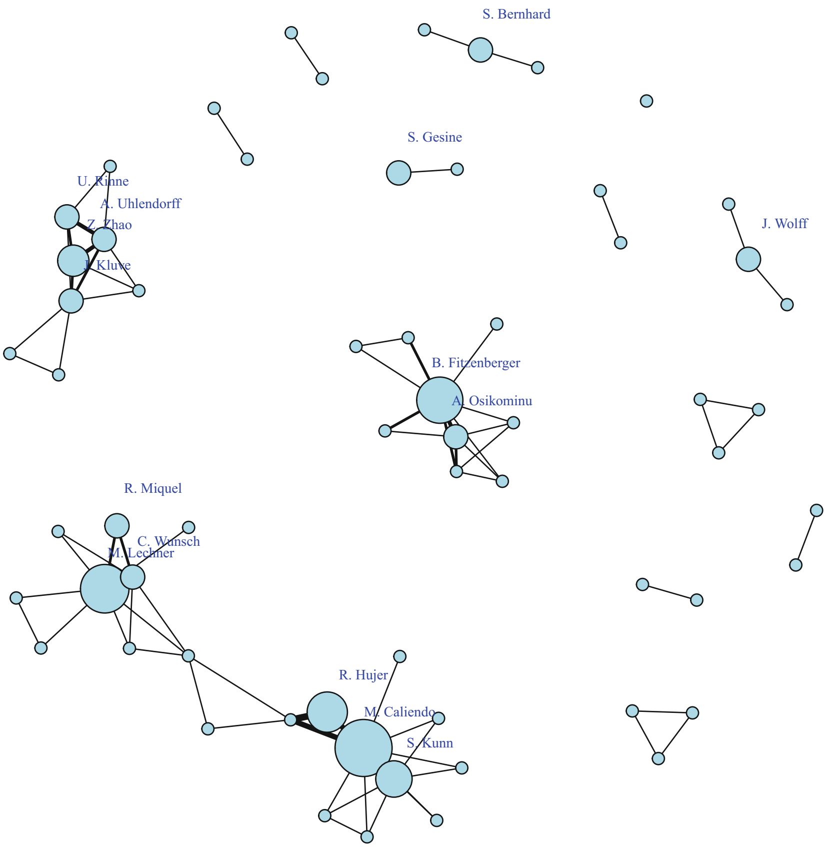


Fig. 3 Network of counterfactual impact evaluation studies in Germany. Source: https://crie.jrc.ec.europa.eu/CIE_database/cieDatabase.php; authors' calculations. Note: This network graph maps the community of authors' networks, through a "node–link" diagram, where the circular nodes represent the authors of CIE studies on German ALMPs and the linear links represent the relationships given by scientific collaborations. The size of the nodes is a proxy of the number of papers of each author

3.3 Time Patterns

The number of CIE publications has clearly increased in recent years (Fig. 4), with 91 of the 111 CIE studies published in 2007 or later. While CIE studies could be identified in 12 countries before 2007, this number increases to 19 when the most recent period is also taken into account. In particular, CIE studies of interventions based in Bulgaria, Ireland, Portugal, Slovakia and Slovenia are observed for the first time in the second period.

The surge of CIE studies in the recent period is certainly partly driven by the rising demand for evidence-based policy. For instance, in the EU provisions for the

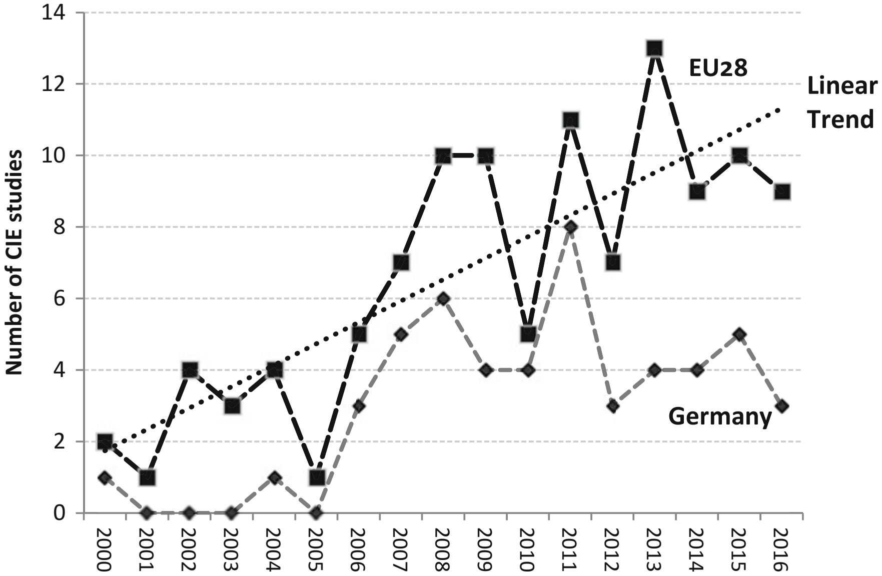


Fig. 4 Counterfactual impact evaluation studies: time patterns. Source: https://crie.jrc.ec.europa.eu/CIE_database/cieDatabase.php; authors' calculations

2014–2020 programming period, impact evaluations have been made compulsory for ALMPs funded through the European Social Fund. This positive time trend in terms of the number of CIE studies is also probably associated with the increasing accessibility of administrative data in several EU countries, as discussed in more detail below.

3.4 Counterfactual Impact Evaluation Methods

As regards the methodology applied in the CIE studies, Table 1 shows that propensity score matching is the approach most commonly employed (54.9%) for evaluating the impact of ALMPs in Europe. The predominance of propensity score matching is even higher if studies based on the combination of propensity score matching with other CIE methods are also taken into account (11.71%). This finding is true for the three categories of ALMPs. Randomised design-based papers rank second (9.91% of all CIE studies). Designs based on a random assignment to the intervention under scrutiny are particularly frequent when it comes to measuring the effect of labour market services (35%). Finally, difference in differences methodology has been implemented for 8.1% of CIE studies.

The pattern observed for the EU-28 is also valid when the summary statistics are limited to Germany. More than half of CIE studies in this country are based

Table 1 Distribution of studies by counterfactual impact evaluation method

CIE method	Frequency	Percentage
Randomisation	11	9.91
Propensity score matching (PSM)	61	54.95
PSM combined with other methods	13	11.71
Difference in differences	9	8.11
Instrumental variables	8	7.21
Regression discontinuity design	4	3.6
Regression/other combinations of methods	5	4.5
Total	**111**	**100**

Source: https://crie.jrc.ec.europa.eu/CIE_database/cieDatabase.php; authors' calculations

on propensity score matching. This method is most frequently applied for the evaluation of training interventions or employment incentives (respectively, 70.6% and 58.8% of these subject-specific studies), while randomisation is more common for measuring the impact of labour market services (63.6%).

3.5 Data Sources

The fact that in Germany the IAB made available to researchers a 2% randomly drawn sample from the integrated employment biographies (IEBs) of the IAB probably largely contributed to promoting CIE-based evidence. The IEBs contain observations on unemployment benefits, job search and participation in ALMPs, combining four data sources.[10] In Nordic countries, such as Finland and Sweden, administrative data have been available to researchers for several years, and hence, unsurprisingly, these countries also rank high in terms of the number of CIEs of ALMPs.

More generally, Fig. 5 reports the distribution of CIE studies by data source. Around 68% of studies are exclusively based on administrative data. The predominance of CIEs based on administrative data is true for the three categories of ALMP, with CIE of training, employment incentives and labour market services based on administrative data in 67.6%, 55.9% and 60% of cases, respectively. The preponderance of administrative data relative to survey data is observed for all CIE methods, though it is more often associated with nonexperimental than with experimental CIE methods (67% and 45.4%, respectively).

As shown in Table 2, propensity score matching is strongly associated with the use of administrative data, with around 67.2% of these studies based on this type of

[10] There are two versions of the IEB samples: a partly anonymised version, which was created in May 2008 and contains data from 1990 to 2008, and a scientific use file, which was built in May 2007 and contains data starting from 1993.

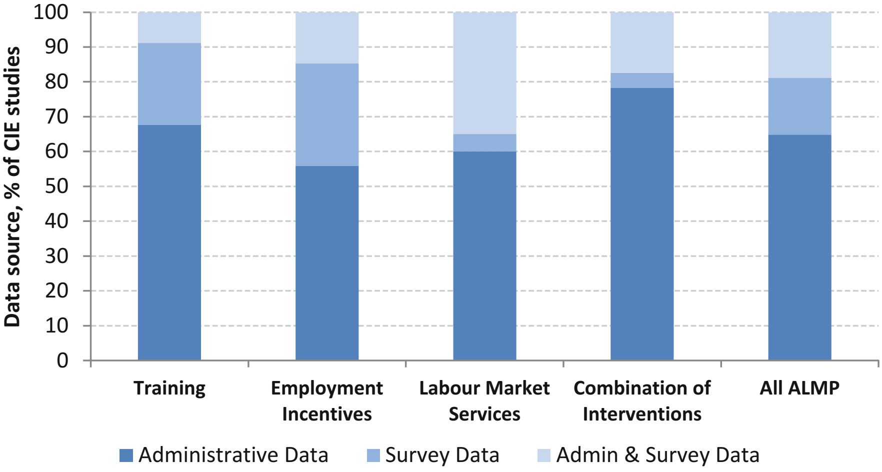


Fig. 5 Counterfactual impact evaluation studies and data sources. (Source: https://crie.jrc.ec.europa.eu/CIE_database/cieDatabase.php, authors' calculations

Table 2 Distribution of studies by counterfactual impact evaluation method and data source

	Data		
CIE method	Administrative (%)	Survey (%)	Combination of data sources (%)
Randomisation	45.45	27.27	27.27
Propensity score matching (PSM)	67.21	11.48	21.31
PSM combined with other methods	53.85	15.38	30.77
Difference in differences	66.67	33.33	0
Instrumental variables	75	25	0
Regression discontinuity design	100	0	0

Source: https://crie.jrc.ec.europa.eu/CIE_database/cieDatabase.php; authors' calculations

data. This figure rises to 88.5% if the CIEs that rely on administrative data merged with survey data are also taken into consideration. CIEs combining propensity score matching with other nonexperimental methods, such as the difference in differences approach, are also largely dependent on administrative data or administrative data merged with survey data. As highlighted in the literature, among others by Sianesi (2002, p. 8) with reference to the evaluation of Swedish ALMPs, the richness of administrative data may justify the use of methods of analysis based on "selection on observables". Indeed, in contrast to an experimental approach (or other nonexperimental methods such as regression discontinuity design), the reliability of the propensity score matching approach for measuring the impact of an intervention critically depends on the validity of the "ignorability" assumption. This CIE approach supposes that the assignment to an ALMP intervention depends only on characteristics (age, previous labour market experience, educational level, etc.) observable by the evaluator.[11] If this is indeed the case then, provided that this selection process is controlled for, variations in the outcome indicators between the participants and non-participants should be due to the participation in the intervention. Along the same lines, the reliability of CIEs employing a difference in differences approach hinges on the availability of longitudinal information (before and after the intervention).

Biewen et al. (2014, p. 838) summarise the importance of data completeness as follows: "for the analysis of ALMP, detailed information on employment and earnings histories prior to program participation seems important to justify matching estimators for treatment effects that rely on a selection on observables assumption. Accurate longitudinal information on labor market transitions is also useful to account for the dynamics of program assignment and to carefully align treated and comparison units in their elapsed unemployment experience".

Administrative data, and preferably a combination of several administrative data sources, allow working on databases containing a large set of information on the participants and non-participants in the interventions. This type of data makes the use of nonexperimental methods for CIE more reliable.

3.6 Administrative Data and Completeness of Counterfactual Impact Evaluation Studies

Although CIE is essential to promoting evidence-based policy, it is also true that not all CIE methods are equally rigorous and informative. The purpose here is not to discuss the assumptions underlying each CIE method but to document whether or not some data characteristics are associated with the comprehensiveness of the impact evaluation. In particular, although it is necessary to document the average effect of an intervention on the participants, it is also of the utmost importance to check if the underlying intervention's impact varies across subgroups of participants. A specific ALMP might not work for the participants as a whole but could be very effective for some subpopulations. If this is not taken into account, the CIE might produce misleading conclusions.

Administrative and survey data differ in terms of population coverage and hence sample size. Indeed, administrative records tend to cover the whole universe of a specific population (for instance, welfare recipients), with this population being tracked for administrative and monitoring purposes, independently of the CIE or any research project. This implies that the sample size used for the CIE of an intervention targeting the population recorded in the administrative database is potentially very large. In contrast, survey data are usually gathered for research purposes, and, as such, sample sizes tend to be much smaller. This is confirmed by the CIE studies in

[11] More specifically, the assignment to the intervention might depend on unobservable characteristics, but these characteristics should not be associated with the outcome of the intervention.

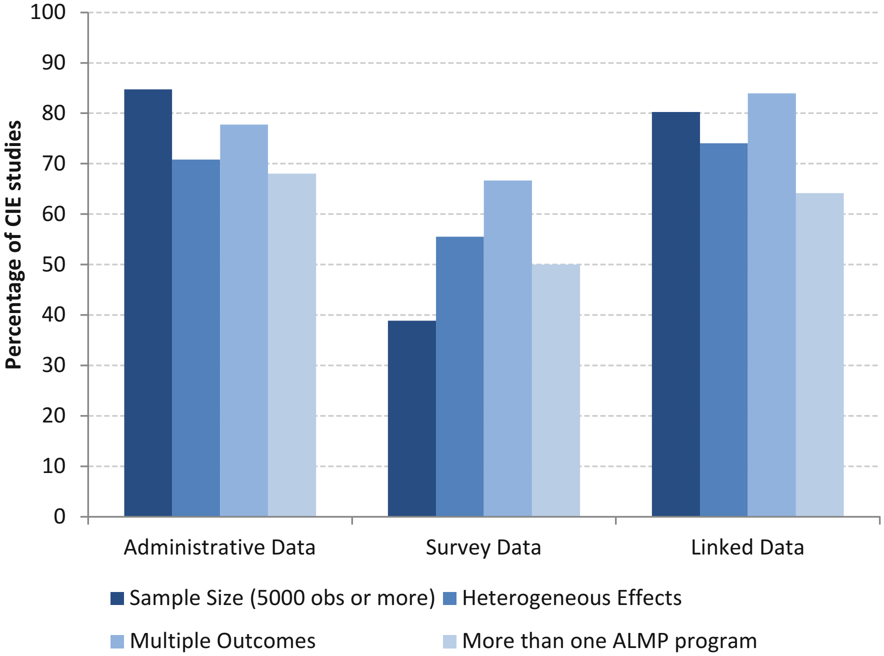


Fig. 6 Data sources and counterfactual impact evaluation completeness. Source: Authors' calculations

the CEA. In almost 85% of CIE studies based on administrative data, the sample size is ≥5000 observations, while this is the case for only 39% of survey-based CIEs. Furthermore, the type of data used in the studies is associated with the likelihood of searching for heterogeneous effects. As shown in Fig. 6, around 72% of CIEs rely on administrative data test for the existence of heterogeneous effects, while this is the case for only 55% of survey-based CIEs.

Along the same lines, CIE studies that examine the effect of ALMPs on short- and long-term outcomes have shown that programme effectiveness can have wide dynamics, from short-term locking-in effects to long-term positive effects on labour market outcomes. In that respect, to ensure a comprehensive CIE, it is important to use a data set allowing for such a dynamic analysis. This is relatively complicated with surveys that are carried out just once or which often suffer from endogenous attrition when repeated over time. In contrast, administrative records generally have a longitudinal component, with the same units being observed repeatedly until being withdrawn from the specific population monitored in the administrative database (e.g. an unemployed person registered in the unemployment office exiting the database when getting hired). As stated in Rehwald et al. (2015, p. 13), "exploring longitudinal register data also allows us to go beyond a single baseline and a single follow-up".

In addition, linking several types of administrative data, such as tax records with employment records, provides the option to examine the effect of a given intervention on a wide range of outcome variables. Among the 93 CIE studies using administrative data, 77.4% examined the effect of the intervention on more than one outcome variable or on the same outcome over a long period of time. This is the case also for the 83.1% of studies that relied on more than one administrative data source or combined administrative data with survey data.

4 Conclusions

The Counterfactual Evaluation Archive (CEA) is an online database, developed by the Centre for Research on Impact Evaluation of the Joint Research Centre of the European Commission, which collects published articles and working papers using counterfactual impact evaluation (CIE) to assess the impact of active labour market policies (ALMPs). This archive assembles information on the studies published over the past 16 years in EU-28 countries.

Using data from this archive, this study observes an unequal distribution of CIEs of ALMPs across EU member states. In particular, while Germany counts 51 CIE studies, some countries, such as Greece, Estonia and Lithuania, are not even represented in the database. Even though there has been an increase in CIE studies in terms of country coverage over recent years, there is clearly still room for improvement. Why is there a lack of evidence for some countries? There are various possible reasons, including a lack of CIE expertise and culture within the country and/or a limitation in data accessibility. In particular, the underutilisation of administrative data, despite the many benefits of such data sources, ranging from the sample size to the longitudinal dimension and the near-universal coverage of the population under study, could explain the still insufficient number of CIEs of ALMPs in many European countries. In Germany, and in some Nordic countries such as Finland and Sweden, administrative data have been available to researchers for several years, and this is probably one of the reasons behind the relatively high number of CIEs of ALMPs observed in these countries.

Analysing the characteristics of the studies included in the CEA, it is argued that CIEs based on administrative data tend to be more comprehensive than those relying on survey data. Administrative data sets have large sample sizes, which facilitate the analysis of heterogeneous effects. Measuring heterogeneous effects is important because a specific intervention might not work for the participants as a whole, but the findings might be different for some subgroups. CIEs based on administrative data or on a combination of sources are more likely to estimate the effect of an ALMP on several outcome variables or to study the short- and long-term impacts of the interventions. Taken together, these statistics suggest that the availability of administrative data is important for promoting evidence-based policy.

References

Ball B, Karrer B, Martin T, Newman MEJ (2013) Coauthorship and citation in scientific publishing. CoRR 1304(0473):abs/1304.0473

Biewen M, Fitzenberger B, Osikominu A, Paul M (2014) The effectiveness of public-sponsored training revisited: the importance of data and methodological choices. J Labor Econ 32(4):837897. http://EconPapers.repec.org/RePEc:ucp:jlabec. https://doi.org/10.1086/677233

Bratu C, Lombardi S, Rodrigues M, Santangelo G, Shaleva A (2014) Knowledge gaps in evaluating labour market and social inclusion policies. European Commission, Joint Research Centre, Ispra

Card D, Kluve J, Weber A (2010) Active labor market policy evaluations: a meta-analysis. Econ J 120(548):452–477

Card D, Kluve J, Weber A (2017) What works? A meta analysis of recent active labor market program evaluations. J Eur Econ Assoc 16(3):894–931

Eberle J, Schmucker A (2015) IZA/IAB Administrative Evaluation Dataset (AED) 1993–2010. FDZ Datenreport, 03(2015)

Jacobi L, Kluve J (2006) Before and after the Hartz reforms: the performance of active labour market policy in Germany. IZA Working Paper No. 2100. Institute for the Study of Labor, Bonn

Kluve J (2010) The effectiveness of European active labor market programs. Labour Econ 16(6):904–918

Newman MEJ (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci USA 98:404–409

Newman MEJ (2004a) Coauthorship networks and patterns of scientific collaboration. Proc Natl Acad Sci USA 101:5200–5205

Newman MEJ (2004b) Who is the best connected scientist? A study of scientific coauthorship networks. In: Ben-Naim E, Frauenfelder H, Toroczkai Z (eds) Complex networks. Springer, Berlin, pp 337–370

Rehwald K, Rosholm M, Svarer M (2015) Are public or private providers of employment services more effective? Evidence from a randomized experiment. IZA Discussion Paper No. 9365. Institute for the Study of Labor, Bonn. http://EconPapers.repec.org/RePEc:iza:izadps:dp9365

Sianesi B (2002) Swedish active labour market programmes in the 1990s: overall effectiveness and differential performance. IFS Paper No. W02/03. Institute for Fiscal Studies, London

Beatrice d'Hombres is a senior scientist at the Joint Research Centre (JRC) of the European Commission since 2006. She holds a PhD in Economics from the Centre for Studies and Research on International Development in France and has more than 10 years of experience in applied micro-econometrics and impact evaluation methods. From 2013 to March 2016, she was the coordinator of the Centre for Research on Impact Evaluation (CRIE) at the JRC. She has been working on issues related to health economics, education, social capital and labour economics. Recent works include the impact of education on attitudes towards immigrants, the effect of smoking ban and health warnings on smoking behaviours. Her research has been published in international journals, including the *European Economic Review*, the *Eastern Economic Journal*, *Health Economics*, the *BE Journal of Economic Analysis* and *Policy or Social Indicators Research*.

Giulia Santangelo is a research fellow at he Joint Research Centre (JRC) of the European Commission, in Ispra (Varese), Italy, since 2013. Since 2017 she is the coordinator of the Centre for Research on Impact Evaluation (CRIE) of the Competence Centre on Microeconomic Evaluation (CC-ME). She obtained a Master of Science in Statistics and a Ph.D. in Economic Sciences at Sapienza University of Rome. She has been a visiting research fellow at Brown University, Rhode Island, USA. Her research interests are in the areas of applied micro-econometrics, labour economics and policy evaluation.

Designing Evaluation of Modern Apprenticeships in Scotland

Matej Bajgar and Chiara Criscuolo

Abbreviations and Acronyms

ABS	Annual Business Survey
APS	Annual Population Survey
ASHE	Annual Survey of Hours and Earnings
CEM	Coarsened exact matching
CSS	Customer Support System
CTS	Corporate Training System
HESA	Higher Education Statistics Agency
HRMC	Her Majesty’s Revenue and Customs
IDBR	Inter-Departmental Business Register
IV	Instrumental variables
LFS	Labour Force Survey
LP	Labour productivity
MA	Modern Apprenticeship
MFP	Multi-factor productivity
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary least squares
PAYE	Pay as you earn
PIAAC	Survey of Adult Skills
RCT	Randomised control trial
SDS	Skills Development Scotland

The opinions expressed and arguments employed herein are those of the authors and do not necessarily reflect the official views of the OECD or of the governments of its member countries.

M. Bajgar (✉) · C. Criscuolo
OECD Directorate for Science, Technology and Innovation, Paris, France
e-mail: matej.bajgar@oecd.org; chiara.criscuolo@oecd.org

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_18

SQA	Scottish Qualifications Authority
UK	United Kingdom
US	United States
VAT	Value-added tax
VET	Vocational education and training

1 Introduction

Apprenticeships combine paid employment and training, with the aim of developing occupational mastery and typically leading to a formal qualification. They have represented an important source of skills since medieval times, and now, in the twenty-first century, they are alive and well. Inspired by the low youth unemployment in those OECD countries where apprenticeships play the largest role—Germany, Austria and Switzerland—more countries are looking to them as a way to provide young people with the skills needed to find fulfilling and well-paid employment.

This is true for the United Kingdom, where apprenticeships are considered an important way of achieving a higher level of skills across the country and a policy worth rigorously evaluating. In England, the government has announced its intention that the number of new apprenticeships offered annually should reach 3 million by 2020 (Department for Business, Innovation and Skills 2015a) and has recently funded several microeconometric studies of the effectiveness of apprenticeship schemes (Bibby et al. 2014; Department for Business, Innovation and Skills 2015b). In Scotland, too, apprenticeships are on the rise, in the form of Modern Apprenticeships (MAs), Foundation Apprenticeships and Graduate-Level Apprenticeships. The number of new MAs available annually in Scotland has increased from 11,000 in 2008/2009 to 25,000 in recent years, and is set to reach 30,000 by 2020. As the livelihoods of more and more young people depend on skills acquired through apprenticeships, and as training subsidies and administrative costs consume an increasing proportion of public funds, Skills Development Scotland (SDS)—the public agency responsible for running MAs in Scotland—has come under mounting pressure to prove that MAs fulfil their goals: to provide young people with the right skills and to increase productivity of Scottish businesses (Audit Scotland 2014).

Up to 2015, SDS evaluated MAs through telephone surveys of employers (Skills Development Scotland 2015) and apprentices (Skills Development Scotland 2013), which asked about the impact on outcomes such as skills, career progression and productivity. Unfortunately, self-report studies are unreliable measures of impact because respondents may consciously or unconsciously adjust their answers to what they expect the evaluator hopes to hear, and simply because it is difficult for them to judge what the outcome would have been without participation. In addition, the surveys contacted apprentices within the first 6 months after leaving the training, and, as a result, they did not provide information on longer term outcomes.

In 2015, SDS decided to examine the impact of MAs through a more rigorous evaluation. SDS's aim was to develop an evaluation that would examine the causal impact of MAs in the long term, could be replicated over time and would allow the effects of MAs on different levels and in different sectors to be compared.

However, it was clear that the evaluation would not be an easy task. The first challenge was a lack of data. Apart from the telephone surveys conducted shortly after the end of training, no information was readily available on outcomes for participating individuals or firms in the longer term. Furthermore, no information had been collected on individuals who had not participated in MAs and could, therefore, serve as a control group. Finally, apprentices had not been selected randomly or through a centralised mechanism, making it hard to separate causal effects of apprenticeships from the effect of characteristics of individuals who had chosen to take up MAs.

Lack of data, absence of a control group and exogenous variation are common limitations of programmes such as apprenticeships. The aim of this chapter is to convince the reader that, despite these drawback situations, although challenging, it is still feasible and valuable to conduct an evaluation that is robust and useful. To this end, it describes the planned evaluation of MAs in Scotland. It builds on a collaboration between SDS and the Organisation for Economic Co-operation and Development (OECD), in which the OECD has prepared an 'evaluation framework' setting out recommendations for evaluating MAs (Bajgar and Criscuolo 2016a).

Designing an evaluation involves answering several central questions. On which outcomes should the evaluator focus? Which data will provide the necessary information while being reliable and accessible? How should the control group be constructed? Which estimation methods should the evaluator use to identify causal effects of the intervention?

Table 1 gives an overview of how some existing studies evaluating impacts of apprenticeships have approached these questions, and describes their findings. It focuses on studies that employ more sophisticated estimation approaches. The table reveals a large diversity of approaches applied by the studies. Many use administrative data but others rely on existing or new surveys. Analysed outcomes include wages, probability of employment, subsequent education and job characteristics. The alternative against which apprenticeships are compared is no training in some studies, but school-based vocational training, unfinished apprenticeships or general education in others. Finally, evaluation methods include the instrumental variables (IV) method, the difference in differences method and randomised control trials.

This chapter describes which evaluation choices are most suitable for evaluating MAs in Scotland, and why. It argues that an encompassing evaluation of MAs should analyse outcomes for both individuals and firms, focusing on employment, wages and several other variables for individuals, and on productivity for firms. *It should rely mainly on existing administrative data,* possibly complemented by a new survey covering information not available in administrative records. The evaluation could use individuals who started but never completed their apprenticeships (non-completers) and those who never underwent an apprenticeship experience (never-starters) as control groups, and conduct estimation over a range of time horizons and

Table 1 Selected studies evaluating impacts of apprenticeships

Authors	Data	Estimation	Findings
Adda et al. (2006)	Social security records	Compare apprenticeships with school-based vocational training; use number of apprenticeship vacancies as IV for participating in apprenticeships	Total wage return 10% after 5 years and 25% after 20 years
Fersterer et al. (2008)	Social security records	Compare completed apprenticeships with unfinished ones; use firm failures as IV for apprenticeship duration	Annual wage return 2–4%
Parey (2008)	Social security records	Compare apprenticeships with school-based vocational training; use number of apprenticeship vacancies as IV for participating in apprenticeships	Annual wage return 3% (not significant with IV); initial reduction in probability of unemployment by 15% points per year of training fades out over time
Malamud and Pop-Eleches (2010)	Census and existing survey	Compare apprenticeships with general education; use regression discontinuity design based on a large educational reform	No causal difference
Alet and Bonnal (2011)	Existing longitudinal education survey	Compare apprenticeships with school-based vocational training; use number of apprenticeship vacancies as IV for participating in apprenticeships	Increase in probability of completing high school diploma and staying in education
Reed et al. (2012)	Programme administrative data and unemployment insurance wage records	Compare apprentices with non-participants and non-achievers; control for initial earnings and use propensity score matching	Participation increases earnings by 50% and the probability of employment by 9% points compared with non-participants; completion increases earnings by 80% and the probability of employment by 15% points compared with non-completers (effects after 6 years)
Picchio and Staffolani (2013)	Administrative data on job contracts	Compare apprenticeships with other types of temporary contracts; use regression discontinuity design based on regional age cut-offs in eligibility	Increase in propensity to get a permanent contract after 2 years

(continued)

Table 1 (continued)

Authors	Data	Estimation	Findings
Bibby et al. (2014)	Linked education, benefit, employment and earning records	Compare apprentice completers to non-completers; use differences estimates with matching for other qualifications	Total wage returns of 11% for Level 2 apprenticeships and of 16% for Level 3 apprenticeships; there is an initial increase in the probability of employment of 3% points but this declines over time
Schaeffer et al. (2014)	Dedicated survey	Compare apprenticeships with attending standard public schools; randomised control trial	No effect on wages; increase in the probability of employment of 26% points; increase in enrolment in General Equivalency Diploma of 24% points; no effect on high school graduation rate
Noelke and Horn (2014)	Labour force survey	Compare apprenticeships with school-based vocational training; use difference-in-differences estimator based on county-level shifts from work-based to school-based vocational education	A 10% increase in the ratio of school- to employer-provided places corresponds to an initial increase in unemployment of 3% points, but this effect declines over time; no effect on working in non-routine occupations
Kugler et al. (2015)	Dedicated survey, and education and social security records	Compare apprentices with non-participants from among preselected candidates; randomised control trial	Total wage return of 6% if formally employed; increase in probability of formal employment of 5% points and of days in formal employment of 13%; increase in probability of completing secondary school of 1.4% points, in the probability of enrolling in college of 3.5% points and in the probability of staying in college after 5 years of 1.6% points

Note: The table is partly based on Table A.1 in Bajgar and Criscuolo (2016b)

using various econometric methodologies. The chapter illustrates how the evaluator can use narrower control groups, matching techniques, regression and, if possible, changes over time to better separate causal effects from mere correlations.

The chapter focuses on apprenticeships, but it also hopes to be informative for evaluations of other types of programmes, and, in particular, of vocational education and training (VET), active labour market policies and education.

Two things should be clarified at this point. Firstly, the chapter does not describe an evaluation that has already been undertaken but an ex ante framework strategy for an evaluation that is just starting. Secondly, this chapter focuses on counterfactual impact evaluation. Other components of programme evaluation, such as input and output monitoring, process evaluation and cost-benefit analysis, are discussed in the OECD evaluation framework for MAs (Bajgar and Criscuolo 2016a), but they are not within the scope of this chapter.

To prepare ground for the subsequent discussion, the following section explains the rationale for public intervention in apprenticeships and provides a brief background on MAs. Section 2 then sets out the scope of the impact evaluation by describing the choice of the units of analysis, the outcomes and the time horizon over which the impact is assessed. Section 3 discusses the choice of data. The next two sections provide guidance on constructing control groups and establishing causal effects for individuals (Sect. 4) and firms (Sect. 5). Section 6 gives illustrative examples of several past apprenticeship evaluations and Sect. 7 concludes.

2 Context

In Scotland, the vast majority of apprenticeships take place within the umbrella of MAs, which are co-ordinated by the government agency SDS. MAs are intended to provide employment opportunities, particularly for young people, and to support economic growth. The aim of MAs is to give young people the skills and the confidence that they need to find a good job and progress in their career. At the same time, they aim to help employers achieve higher productivity and boost staff retention and morale. MAs combine paid work with training that leads to a recognised qualification and are offered at several levels, ranging from Foundation Apprenticeships for secondary-school pupils to Graduate-Level Apprenticeships that are equivalent to a master’s degree.

A feature of MAs that is important for their evaluation is that SDS does not run them as a centralised programme but rather provides standards, oversight and financial contributions to many privately run programmes under the MA umbrella. A key role, therefore, is played by training providers (e.g. private training providers and colleges), which deliver learning and assessment. SDS allocates apprenticeship places to the training providers, and provides a contribution to the training fees, although in some cases employers are also asked to contribute to the training costs.

3 Scope of Evaluation

This section demarcates the scope of the evaluation in terms of units of analysis, examined outcomes and time horizon.

3.1 Units of Observation

The first question to ask when designing an impact evaluation is 'impact on whom'. The answer to this question for MAs is that it should examine the impact on individuals and on firms.

Each type of unit has distinct advantages for measuring the impact of apprenticeships on economic growth. The key advantage of using individuals as the unit is that, for a young person, participating in an apprenticeship potentially represents one of the most important determinants of his or her labour market performance. This will imply a strong 'signal-to-noise ratio' which in turn allows for a more precise estimation of MAs' impact on apprentices' outcomes. The key advantage of using firms is that it is easier to measure productivity in a comparable way than it is when using individuals. However, apprentices represent only a small percentage of the workforce for most employers; thus, the effect of the apprenticeships on firm productivity may be difficult to discern even if the effect on any single apprentice is large.

3.2 Outcomes

The second key question in establishing the scope of the evaluation is 'impact on what': That is, which outcomes should it examine? As illustrated in Table 1, the answers to this question given by existing studies are most commonly wages and employment probability, but in some cases also job characteristics and subsequent education.

As the purpose of MAs is to develop young people's skills and productivity, these two outcomes appear to be the most obvious options when evaluating the impact of MAs on individuals. Unfortunately, measuring either of them in a comparable way is challenging.

Skills are not readily observed in existing data and would need to be newly measured in a costly and methodologically demanding survey. Furthermore, the applied nature of apprenticeship training means that a large part of the acquired skills is industry or occupation specific. Therefore, measures of such skills cannot easily be compared across industries or occupations. This is an important limitation, because comparing performance of different MA frameworks is one of the main aims of the evaluation as seen by SDS.

Direct measures of workers' individual productivity are similarly occupation specific and thus hard to compare across occupations, and even across tasks within a given occupation. In addition, records of such measures may exist within individual employers but are unlikely to exist for a representative set of employers.

For these reasons, evaluation of MAs should instead focus on several other individual-level outcomes:

- Employment
- Wages
- Unemployment
- Standard/non-standard employment[1]
- Career progression
- Subsequent education
- Subjective well-being

In the case of impact on employers, productivity plays a central role. MAs aim to support economic growth, and increasing firm productivity is the main way in which they can do so. Unlike individual-level productivity, firm productivity can be measured in a comparable way, most commonly as one of the following two measures:

- **Labour productivity (LP)** can be defined as output per hour worked and can be calculated as the ratio of sales or value added over the number of employees, measured as head counts or preferably, if the information is available, as full-time equivalents.
- **Multi-factor productivity (MFP)** captures the output a firm can produce with a given amount of inputs (e.g. labour, capital, total intermediate inputs, energy). It differs from LP in that it also accounts for the amount of physical capital and intermediates used by each firm. The disadvantage of MFP is that it requires more information on inputs, such as physical capital and intermediates.

3.3 Time Horizon

The third question for the evaluation is ‘impact, but when’.

This question is important because existing evaluations suggest that estimated effects vary substantially depending on the time between the end of the apprenticeships and the moment when the impact is assessed. The effect on employment is often large immediately after training but gradually declines over the next few years.[2] In contrast, some studies suggest that the effect on wages grows substantially over time.[3] Mechanisms for the effects may differ over time, too. While short-term employment effects are probably due to apprentices staying with the same employer shortly after the apprenticeship ends, the later wage increases may instead reflect acquired skills.

For individuals, it seems optimal to examine the outcomes over multiple time horizons. Such an approach will provide richer information on the impact and the mechanism behind each outcome. This approach, however, requires the use of large

[1] Standard employment here stands for full-time employment with an open-ended contract.

[2] See Bonnal et al. (2002), Parey (2008), Noelke and Horn (2014) and Bibby et al. (2014).

[3] See Cooke (2003) and Adda et al. (2006).

administrative data sources that can be accessed at different points in time and still contain information for a large number of apprentices. With survey data, a particular time horizon would need to be chosen; 2–4 years after the end of the MAs may provide a sufficient time for the effects to materialise while still allowing a reasonably precise identification of the effects.

In the case of impact on firms, the question of time horizon concerns the time between the training of the apprentices employed by a firm and the time when the outcomes are measured. While the analysis on firms, too, would ideally allow for a different effect of apprentices depending on the time since training, in reality the number of apprentices employed by most firms is small and identifying such a differential effect may not be possible.

4 Data

An essential part of any evaluation is the choice of data. In the case of individuals, the evaluation requires information on whether or not he or she took part in MAs, on the examined outcomes and on individual or firm characteristics. The first three parts of this section discuss three main ways of obtaining all this information: using a self-standing existing survey, linking multiple data (from administrative sources or from surveys) and collecting new information via a new survey specifically designed and conducted for evaluating MAs. The fourth part discusses the choice of data for evaluating the effect of apprenticeships on employers.

4.1 Data on Individuals: Existing Self-Standing Surveys

The simplest option is to rely on a single existing survey. In the Scottish context, surveys which could be considered include the Labour Force Survey (LFS) and the Annual Population Survey (APS).

The LFS gathers quarterly information on demographics, education and labour market outcomes of 100,000 individuals living in the United Kingdom. It includes a variable describing if an individual has completed an apprenticeship, but it does not specifically refer to MAs. The APS is obtained by merging waves 1 and 5 of the LFS quarters with annual boosts for England, Scotland and Wales. It samples about 340,000 individuals each year, although it contains fewer variables than the standard LFS. It also has a panel element: the households at selected addresses are interviewed annually over four waves and then leave the survey and are replaced by other households.

Using a single existing survey is a cheap and fast option, because it taps into information that already exists and is in one place. However, the number of Modern Apprentices in the surveys is rather small, the surveys cannot be linked to SDS individual records and they do not identify individuals who started but did not complete their apprenticeships. These features limit the potential precision of estimates, complicate disaggregated analysis and limit options for constructing control groups.

For these reasons, existing self-standing surveys are not a sound basis for a robust evaluation of the impact of MAs on individuals.

4.2 Data on Individuals: Linked Administrative and Survey Data

An alternative approach, one which is taken by most studies described in Table 1, is to combine information from several sources. In the case of MAs, the main potential sources of information include programme records held by SDS, Her Majesty's Revenue and Customs (HMRC) employment and earnings records, Department for Work and Pensions benefit histories, Annual Survey of Hours and Earnings (ASHE), educational records and the business register.[4]

Programme records held by SDS include the Corporate Training System (CTS) and the Customer Support System (CSS). CTS contains complete information about each apprentice's training. CSS contains information about education, training and labour market participation of 16- to 19-year-old cohorts, including young people who did not enter an apprenticeship and could thus serve as a control group.

The relevant administrative records held by HMRC include the P45 employment data, which include information on employment start and end dates, and on whether or not individuals are claiming benefits, and the P14 earnings data, which include information on employment start and end dates, earnings and hours worked. This information can be complemented with the records held by the Department for Work and Pensions, which contain detailed information on benefits claimed by each individual.

ASHE is a survey filled in by employers, and provides job and earnings information on approximately 140,000–185,000 individuals each year. It covers only a fraction of the population, but, in addition to employment and wage figures, it specifies if employees have a permanent contract and if they have a managerial role; as a result, it allows additional outcomes, such as career progression and employment stability, which are not covered by the HMRC data, to be examined.

Administrative records on education—the Pupil Census, Scottish Qualifications Authority (SQA) Attainment Data and the Higher Education Statistics Agency (HESA) Student Record—provide information on individuals' schooling and their qualifications other than MAs. The Pupil Census contains mostly demographic

[4] A similar approach, linking further education learner information with data from HMRC and the Department for Work and Pensions, has been used for evaluating outcomes of further education in England (Bibby et al. 2014).

information on all pupils in publicly funded schools in Scotland. The SQA Attainment Data contain information on entry and attainment of SQA qualifications. The HESA Student Record is an administrative dataset of all students in higher education in Scotland.

Finally, the Inter-Departmental Business Register (IDBR) can provide complementary information on employers with which apprentices undertook their training, and on their current employers. It is a comprehensive list of UK businesses based mainly on value-added tax (VAT) and pay-as-you-earn (PAYE) systems, both administered by HMRC, but also on information from Companies House, Dun & Bradstreet and business surveys. The information covered includes industry, employment, turnover and country of ownership.

Linking the datasets requires *matching records* pertaining to each person. This is preferably done based on a unique individual identifier. A more laborious and less precise method is to rely on personal information such as name, address and date of birth, and should be used if only the first option is not feasible. Since the SDS records on Modern Apprentices (CTS) include the National Insurance number, they can be linked to HMRC and ASHE data. The records on other SDS programmes (CSS) can, in turn, be linked to the educational data using the Scottish Candidate Number, which is allocated to school pupils sitting SQA exams. A link between CTS and CSS records can then connect the data linked with the National Insurance number with those linked with the Scottish Candidate Number. Finally, employer information can be matched to the individual-level data using the IDBR number and the PAYE number.

The data linking approach has some important advantages. In particular, the very large sample size of the resulting dataset allows obtaining precise estimates, decomposing results by MA types or individual characteristics, defining narrower control groups better matching people who undertook MAs, and examining outcomes at different horizons. The approach also provides information that does not appear in existing surveys, for example type and exact dates of MA or information on career progression. Moreover, once the data are set up, the marginal costs of adding new observations and re-running the estimation are relatively low.

Setting up the linked dataset requires a significant initial sunk investment in time and effort to access, link and clean the data. However, once the data infrastructure is in place, it will provide a powerful tool for evaluating not only MAs but also other public interventions in Scotland. It should, therefore, be used as the main information basis for the impact evaluation of MAs.

4.3 Data on Individuals: Dedicated Survey

The third option is to conduct a new survey specifically for the purpose of evaluating the impact of MAs. This is the choice made by studies analysing randomised control trials, such as Schaeffer et al. (2014) and Kugler et al. (2015), although the latter study combines newly collected survey data with administrative datasets to look at longer term outcomes.

The advantage of this option is that it *gives the greatest freedom in collecting the exact information* that will be useful for the evaluation. On the other hand, it is the most costly option, particularly when used for collecting all the information that is needed for evaluation purposes and in case of repeated evaluations. It also suffers from the same disadvantages as using existing surveys: limited sample size and wage information that is less reliable than administrative data, since the data are self-reported.

The advantages and disadvantages of a dedicated survey mean that it is not the best option for evaluating the impact on most core outcomes, such as employment and wages. It could, nevertheless, be beneficially used for evaluating the impact of MAs on outcomes not captured in administrative data, most notably subjective well-being and skills. To maximise value for money, the survey should be undertaken only once in several years and its results should be matched with the linked administrative and survey data.

4.4 Data on Employers

Since no existing surveys capture which firms participate in MAs, evaluating the impact of MAs on employers requires either linking employer records with information held by SDS or conducting a new survey.

The information needed to calculate productivity is available from the IDBR and from the Annual Business Survey (ABS). The IDBR covers the whole firm population but allows calculation of labour productivity only measured as turnover over employees. The ABS, in contrast, covers only a sample of firms but contains information on value added and investment.

To obtain information on participation in MAs, these datasets need to be linked to the CTS records held by the SDS. An experimental link between CTS and IDBR has already been established based on employer names and post codes. ABS data can then be added based on the IDBR number. In the future, collecting employer identifiers—IDBR, PAYE, VAT and company registration number—by SDS would allow a simpler, more comprehensive and more accurate link.

As with individual-level data, an alternative approach is to conduct a new survey. However, as apprentices represent only a small proportion of employees for most employers, accurately measuring their effect in a survey of a limited size is challenging. For this reason, the individual-level survey should be given priority or the employer survey should focus on small- and medium-sized enterprises (SMEs), where both benefits and costs of MAs may be easier to measure and which are harder to observe in the ABS.

5 Counterfactual Impact Evaluation (Individuals)

Impact evaluation aims to answer this question: ‘What would have happened to the person/firm had the intervention not taken place?’ The central challenge lies in the fact that, at a given point in time, it is not possible to observe the same person or firm both with and without the evaluated intervention (e.g. apprenticeship). The hypothetical alternative ‘counterfactual’ scenario is fundamentally unobservable.

For this reason, impact evaluation relies on ‘control groups’. These consist of units (e.g. individuals, firms) which have not been exposed to the intervention but were initially as similar as possible to units which have been exposed to the treatment, i.e. the ‘treatment group’.

As it is difficult to find a control group that is the same as the treatment group in all respects other than participation in the evaluated intervention, a range of approaches can be used for making the treatment and control groups as similar as possible, and for taking into account any remaining differences between them.

This section discusses the choice of control groups and evaluation methods for assessing the impact of MAs on individuals. The subsequent section will then focus on the impact on employers.

5.1 Control Groups

There are, in principle, two groups of people from which the control group for the evaluation of MAs could be drawn. One group are individuals who never started an MA—‘never-starters’. The other group are people who started an MA but did not complete it—‘non-completers’.

The use of each of these two groups has different advantages and challenges. The main challenge in the case of never-starters is that starting an MA is not random and may be related to a person’s characteristics, such as skills, motivation, socio-economic background and local economic conditions. These characteristics may, in turn, influence outcomes such as employment, wage and subjective well-being. Consequently, observed differences in these outcomes between never-starters and MAs may be due not only to a causal effect of taking an MA but also to differences in individual characteristics.

The problem of non-random selection into MAs does not apply to non-completers because they have themselves started an MA. Instead, one challenge with non-completers is due to a non-random selection into non-completion. Non-completers may lack skills and determination, or non-completion may be a result of more attractive outside opportunities. In either case, differences in outcomes could be due to differences in personal characteristics rather than the causal effect of completing an MA. An additional challenge with using non-completers as a control group is that apprentices are likely to learn from their MA even if they do not complete it. The difference in outcomes between MA non-completers and completers, therefore, may capture only part of the full benefit of MAs and thus underestimate the positive impact of apprentices on individuals.

The preferred solution is to use both types of control groups—non-completers as well as never-starters—for the analysis. While both control groups face problems due to non-random selection, the nature of the selection is different in each case. Finding that both approaches lead to similar results will be reassuring; if the results differ, the direction of the difference will be informative for interpretation of the estimates. Furthermore, using multiple control groups is not particularly costly when administrative data are used for evaluation. However, it may not be a suitable choice for evaluation based on newly collected survey data.

An important additional decision is how to account for other qualifications that young people in the treatment and control group achieve. Doing so is important for separating the causal effect of MAs, on outcomes such as employment and wages, from the effect of other qualifications. Restricting the treatment and control groups to individuals with no higher qualifications should be preferred in the baseline specification, and using the full sample and controlling for other qualifications would be a useful complementary analysis, providing additional information.

5.2 Evaluation Methods

The aim of the approaches discussed here is to separate the causal impact of Modern Apprenticeships from mere correlations that may be due to differences in individual characteristics between the treatment group and the control groups. While some popular approaches are not feasible in the context of Modern Apprenticeships, several others can significantly improve the reliability of the estimated impact.

A method that is widely regarded as the ‘gold standard’ in impact evaluation is *randomised control trials* (RCTs). RCTs randomly divide individuals into a treatment group and a control group, thus overcoming the non-random selection issues discussed above. Unfortunately, applying this approach to the evaluation of MAs has not proven to be possible.[5]

An alternative to conducting an experiment would be to rely on factors that make apprentices more likely to participate in MAs but do not directly affect their labour market outcomes. Such ‘instrumental variables’ could be related, for example, to the availability of MA places in a given location and year,[6] to regional differences

[5] For an evaluation of an apprenticeship programme using a RCT, see Schaeffer et al. (2014).

[6] For examples, see Adda et al. (2006), Parey (2008), Alet and Bonnal (2011) and Noelke and Horn (2014).

in relevant regulations[7] or to the intensity with which MAs are promoted in different regions or at different schools. Alternatively, the analysis could rely on factors that cut some apprenticeships short for reasons that are random from the apprentice's perspective.[8] Unfortunately, no suitable source of exogenous variation for which information is available has been identified in the context of MAs.

Several other approaches can instead be used to conduct the evaluation. To begin with, the control groups can be defined more narrowly, in order to more closely match the treatment group.

Firstly, the risk that never-starters are systematically different from apprentices in their characteristics, and that this drives the observed outcomes, can be partly overcome by restricting the control group to individuals to whom an SDS career counsellor has suggested entering MAs. This information is available in the CSS records and should be used to strengthen the estimation.

Secondly, the selection problem in the case of non-completers, which arises because the factors underlying non-completion may also affect later outcomes, can be addressed using information on the reason for non-completion, available in the CTS records held by SDS. One option is to focus only on those individuals who failed to complete their MA for a reason that is unlikely to affect later labour market outcomes (e.g. moving to a new location).[9] An alternative is to split non-completers into those whose reason for non-completion is likely to be negatively related to later outcomes (e.g. lack of motivation) and those whose reason for non-completion is likely to be positively related to them (e.g. being offered a different job). As the first control group would likely lead to overestimation of the effect of MAs and the latter control group to its underestimation, they would provide, respectively, an upper and lower bound for the estimate.

Thirdly, the issue that even incomplete training is likely to produce some benefits can be largely addressed by restricting the control group to those non-completers who left their training shortly after it began.

Matching techniques are another way of making the control group as similar to the Modern Apprentices as possible. In particular, they can be used to construct a control group (based on never-starters or non-completers) with individuals who are similar to apprentices in terms of their *observable* characteristics, such as gender, age, socio-economic background, previous education, region and previous labour market history. Based on these characteristics, apprenticeship completers can be matched with members of the control group who were initially similarly likely to start or complete an apprenticeship (using a different version of propensity

[7] See Picchio and Staffolani (2013).

[8] See Fersterer et al. (2008).

[9] A similar strategy has been used to evaluate private on-the-job training (Leuven and Oosterbeek 2008; Görlitz 2011).

score matching)[10] or whose initial characteristics were similar to those of the apprenticeship completers (coarsened exact matching (CEM)).[11]

Defining a suitable control group is essential for the evaluation and regression methods should not be seen as its substitute. But once an appropriate control group is constructed, regression analysis can be beneficially used to control for further factors that could distort the estimated relationship between participating in an MA and the examined outcomes. Such factors may include previous education and the region, size and industry of the MA employer.

Matching and regression techniques usefully account for observable individual characteristics but cannot account for the unobservable ones. A complementary approach could therefore be to analyse changes in outcomes for given individuals over time, rather than focusing on levels. This would allow controlling for individual characteristics which are constant over time and hard to directly observe in the data (e.g. motivation, talent or persistence) but which may be important in determining the evaluated outcomes. For each individual, the approach would compute changes in outcomes (e.g. wages) between the periods before and after the apprenticeships (or the time when individuals in a control group were most likely to start an apprenticeship), and then it would compare these differences between the apprentices and the control group. Importantly, this desirable approach is subject to the availability of a sufficient number of apprentices with work experience prior to their apprenticeship.[12]

6 Counterfactual Impact Evaluation (Employers)

Evaluation of impact on employers aims to compare the actual performance (e.g. productivity) of firms participating in MAs with the performance that the employer would have achieved had the firm not participated in MAs, or had it participated to a different extent. As in the case of individuals, the 'counterfactual' scenarios cannot

[10] Introduced by Rosenbaum and Rubin (1983), propensity score matching proceeds in two steps. The first estimates the probability of being treated (e.g. participating in an apprenticeship) based on observed individual characteristics. The second step then matches individuals with similar predicted probability—propensity score—from the estimation. For example, each individual in the treatment group can be matched with the person in the control group with the most similar propensity score ('nearest-neighbour matching') or can be compared with a linear combination of multiple individuals, with weights given by differences in the propensity score. For applications in the context of on-the-job training, see Almeida and Faria (2014) and Kaplan et al. (2015), and for an application to evaluating apprenticeships see Reed et al. (2012).

[11] CEM is a more robust way than PSM to construct a control group that is actually similar to the treatment group. However, CEM requires a very large sample size, and its successful application would, therefore, require comprehensive administrative data. CEM is described by Iacus et al. (2011) and has been used in the context of apprenticeship evaluation by Bibby et al. (2014).

[12] Reed et al. (2012) use changes over time to evaluate apprenticeships, and Bibby et al. (2014) do so when evaluating further education.

be directly observed and have to be approximated by other, otherwise similar, firms which differ only in terms of their participation in MAs.

This section discusses how this could be done. First, it proposes a way in which the intensity of firm participation in apprenticeships could be measured. Then it discusses evaluation approaches that can be used to strengthen the estimation.

6.1 Measuring Intensity of Participation in Modern Apprenticeships

For firms, unlike individuals, participating in MAs is not a binary decision. In addition to the question of whether or not a firm participates at all, it also matters how many apprentices it takes on relative to its size.

A recent review of literature on apprenticeship evaluations (Bajgar and Criscuolo 2016b) suggests that employing a regular worker without training is the most common alternative to training an apprentice. For this reason, the number of Modern Apprentices relative to the total number of workers employed by a firm is the most appropriate measure of firm participation in MAs.

It is also important to consider whether the number of Modern Apprentices should be based on current or past apprentices, because the effect of apprentices in training, relative to regular workers, is likely to be negative even if the long-term effects are positive and large. To measure long-term effects that operate through apprentices' individual productivity, the evaluation should focus on the effect of employing past Modern Apprentices. In contrast, a potential evaluation of the short-term effects of training Modern Apprentices (e.g. the fact that MAs spend less time than regular employees actually working, boosted staff morale) should focus on the number of apprentices currently in training.

6.2 Evaluation Methods

Evaluating the impact of MAs on employers also requires the causal effect of MAs to be separated from other factors that may be correlated with both firms' involvement in MAs and the examined outcomes. On one hand, firms that take on apprentices might be better managed. Such firms would be more productive, but not as a result of their participation in MAs. On the other hand, some firms may train apprentices in response to underinvestment in training in previous years. They would probably have lower productivity, but again not as a consequence of training apprentices.

Several methods can be employed to bring the estimated impact closer to the true value. An RCT with an 'encouragement design' could be used to induce randomly selected firms to participate (e.g. through a training voucher). Such an experiment would provide exogenous variation for identifying the true impact of MAs, and, in addition, it would provide valuable information on the extent to which financial incentives help stimulate firms' interest in MAs. However, a similar experiment would require strong policy support and would take time to allow the measurement of long-term outcomes, and as a result will not be part of the initial impact evaluation of MAs.

Instrumental variable techniques would exploit factors which increase some firms' participation in MAs without affecting firm performance other than through the apprenticeships. Such factors could be related to MA contribution rates or varying institutional arrangements for MAs across different occupations or regions.[13] Unfortunately, there do not seem to be any observed variables that could be used in this role in the Scottish context.

Instead, matching techniques can be used to construct a control group of firms that do not engage in MAs, but which are similar to the firms that do in their other characteristics. The matching could be based on, for instance, industry, size, age and availability of a training budget.

In addition, regression analysis can be used to take into account observed firm characteristics that could be related both to participation in MAs and to the examined outcomes, such as size and training intensity.

For most firms, unlike for many individuals, the examined outcomes can be observed both before and after participating in MAs. The evaluation should compare changes in examined outcomes over time between firms that start taking on apprentices and those that do not and, by doing so, account for unobservable time-invariant firm characteristics which could otherwise bias the results.

7 Examples of Past Evaluations

This chapter describes a design of an evaluation which has yet to be undertaken. In this section, the evaluation design is illustrated by briefly describing several apprenticeship evaluations which have already been conducted and have produced interesting results. The first two studies, from England and from the United States, are included because they use data and methods somewhat similar to those proposed for evaluating MAs in Scotland. Two other studies, both evaluating the same programme in Colombia, then give an example of an RCT in the context of apprenticeships.

7.1 Impacts of Apprenticeships in England

The study by Bibby et al. (2014) is among a series of studies conducted by the UK Department of Business, Innovation and Skills to evaluate effects of further

[13] Cappellari et al. (2012) examine the effect of apprenticeships on firm productivity using random staggering of a policy change roll-out across Italian regions and industries.

education in England. It also examines other types of further education, but it includes a specific section on apprenticeships. It relies on a linked administrative dataset, which this chapter also recommends for evaluating MAs in Scotland. This dataset includes individual education histories, information on benefits from the Department for Work and Pensions, and employment and earning information from HMRC.

To estimate the effects of apprenticeships, the study compares outcomes for apprenticeship completers with those for individuals who start apprenticeships but do not complete them. It estimates the effects by ordinary least squares (OLS), controlling for a number of individual characteristics such as gender, age, disability, region and prior education.

The study finds a daily wage premium associated with completing an apprenticeship of 11% and 16% for Level 2 and Level 3 apprenticeships, respectively. It also finds an employment probability premium of almost 3% points shortly after the end of the training, which nevertheless disappears by year 4 or 5. It also shows that its results are reasonably robust compared with several more sophisticated estimation approaches, although it pursues these only for further education overall and not specifically for apprenticeships. It applies difference-in-differences or difference-in-differences-in-differences methods, combined with CEM (Iacus et al. 2011), and it restricts the control group to individuals who dropped out of their training early on, finding that this leads to a larger estimated effect of the further education qualifications.

7.2 Impacts of the Registered Apprenticeships in the United States

Reed et al. (2012) estimate the impacts of the Registered Apprenticeships in ten US states using programme administrative data and unemployment insurance wage records. They estimate a regression comparing apprentices who completed different proportions of their apprenticeships, controlling for their initial earnings. Alternatively, they also compare completers and non-completers using propensity score matching, using the initial earnings as one of the variables for estimating the propensity score.

The results suggest that completing a Registered Apprenticeship is associated with annual earnings that are, on average, approximately USD 6600 higher in the sixth year after the start of the training, with the earnings premium dropping only slightly, to USD 5800, in the ninth year.

7.3 Impacts of the Youth in Action Programme in Colombia

Three studies analyse the impacts of the Youth in Action programme, which was in place in Colombia between 2002 and 2005. The programme was targeted at poor young people living in cities, and it involved a 3-month classroom-based training programme followed by a 3-month apprenticeship in a company operating in the formal sector of the economy. The studies benefit from the fact that, in 2005, the Colombian Government agreed to conduct an experiment in which it allocated the oversubscribed training places among preselected candidates based on a lottery. The evaluations can thus rely on a control group consisting of individuals who were interested in the training, and well qualified to enter it, but who were not given the opportunity to participate.

Attanasio et al. (2011) studied the short-term effects of the programme using a baseline survey carried out shortly before the start of the training and a follow-up survey conducted 13–15 months after its end. Their results suggest that the programme increased female employment by 7% points and female total earnings by 20%. The estimated increase in male employment and earnings was much smaller and not statistically significant. Both women and men were 5–7% more likely to be employed in the formal sector of the economy as a result of the programme.

Kugler et al. (2015) explored the long-term effects of the programme using administrative social security and education records matched to information on programme participants and control group. They found that even 3–8 years after the experiment the programme participants were 5% points more likely to be in formal employment; had spent, on average, 13% more days in formal employment; and, if they worked in the formal sector, their daily wage was 6% higher compared to members of the control group who also work in the formal sector. They were also 1.4% points more likely than the control group to have completed secondary school, 3.5% points more likely to have enrolled in college and 1.6% points more likely to have stayed in higher education 5 years after the training.

8 Concluding Remarks

This chapter has described the planned impact evaluation of MAs in Scotland. It has focused on the challenges the evaluation faces and on reasons for taking some approaches over others. It has aimed to encourage and inform similar evaluations elsewhere. With this aim in mind, this closing section highlights three interlinked and more general lessons learned from the evaluation strategy for MAs.

The first lesson emphasises the *potential of using administrative data* for similar evaluations. The advantages of administrative data, and especially of the large coverage they offer, have been reiterated throughout the chapter: they are instrumental for estimating effects separately for different types of MAs and individuals with different characteristics; they allow us to analyse how the effects of MAs evolve over time after the end of the training; they make it possible to use multiple control groups; and they facilitate identification of causal effects through the use of narrower control groups and matching techniques.

Second, setting up a linked dataset and conducting the evaluation can be greatly facilitated by collecting the right information in programme administrative records early on. In the case of MAs in Scotland, this would involve collecting relevant individual and company identification numbers and ensuring that the collected variables that are important for the evaluation (e.g. reason for non-completion) are coded, complete and correct.

The last lesson underlines *the importance of having an ex ante strategy* for ex post programme evaluation. Such a strategy highlights the data and variable requirements of the evaluation and, by doing so, allows early collection of the necessary information, as discussed in the second lesson. In addition, the strategy may encourage designing the programme ex ante in a way that allows robust evaluation approaches based on RCTs, regression discontinuity design or instrumental variables ex post.

Acknowledgements The authors would like to thank the staff of Skills Development Scotland for contributing important ideas and information to the analysis.

References

Adda JC, Dustmann C, Meghir C et al (2006) Career progression and formal versus on-the-job training. IZA Discussion Paper No 2260. Institute for the Study of Labor, Bonn

Alet E, Bonnal, L (2011) Vocational schooling and educational success: comparing apprenticeship to full-time vocational high-school. Toulouse School of Economics, Toulouse

Almeida RK, Faria M (2014) The wage returns to on-the-job training: evidence from matched employer-employee data. IZA Lab Dev 3:1–33

Attanasio O, Kugler A, Meghir C (2011) Subsidizing vocational training for disadvantaged youth in Colombia: evidence from a randomized trial. Am Econ J Appl Econ 3:188–220

Audit Scotland (2014) Modern apprenticeships. Audit Scotland, Edinburgh

Bajgar M, Criscuolo C (2016a) OECD evaluation framework for modern apprenticeships in Scotland. OECD Science, Technology and Industry Policy Paper 2016/35. Organisation for Economic Co-operation and Development, Paris

Bajgar M, Criscuolo C (2016b) Impact of apprenticeships on individuals and firms: lessons for evaluating modern apprenticeships in Scotland. OECD Science, Technology and Industry Working Paper 2016/06. Organisation for Economic Co-operation and Development, Paris

Bibby DF, Buscha A, Cerqua D et al (2014) Estimation of the labour market returns to qualifications gained in English further education. Research Paper 195. Department for Business, Innovation and Skills, London

Bonnal L, Mendes S, Sofer C (2002) School-to-work transition: apprenticeship versus vocational school in France. Int J Manpow 23:426–442

Cappellari L, Dell'Aringa C, Leonardi M (2012) Temporary employment, job flows and productivity: a tale of two reforms. Econ J 122:188–215

Cooke LP (2003) A comparison of initial and early life course earnings of the German secondary education and training system. Econ Educ Rev 22:79–88

Department for Business, Innovation and Skills (2015a) English apprenticeships: our 2020 vision. HM Government, London

Department for Business, Innovation and Skills (2015b) Measuring the net present value of further education in England. BIS Research Paper No 228. BIS Research, London

Fersterer J, Pischke J-S, Winter-Ebmer R (2008) Returns to apprenticeship training in Austria: evidence from failed firms. Scand J Econ 110:733–753

Görlitz K (2011) Continuous training and wages: an empirical analysis using a comparison-group approach. Econ Educ Rev 30:691–701

Iacus SM, King G, Porro G (2011) Multivariate matching methods that are monotonic imbalance bounding. J Am Stat Assoc 106:345–361

Kaplan DS, Novella R, Rucci G et al (2015) Training vouchers and labor market outcomes in Chile. Working Paper 585. Inter-American Development Bank, Washington, DC

Kugler A, Kugler M, Saavedra J et al (2015) Long-term direct and spillover effects of job training: experimental evidence from Colombia. NBER Working Paper 21607. National Bureau of Economic Research, Cambridge

Leuven E, Oosterbeek H (2008) An alternative approach to estimate the wage returns to private-sector training. J Appl Econ 23:423–434

Malamud O, Pop-Eleches C (2010) General education versus vocational training: evidence from an economy in transition. Rev Econ Stat 92:43–60

Noelke C, Horn D (2014) Social transformation and the transition from vocational education to work in Hungary: a differences-in-differences approach. Eur Sociol Rev 30:431–443

Parey M (2008) Vocational schooling versus apprenticeship training – evidence from vacancy data. Institute for Fiscal Studies. http://cep.lse.ac.uk/seminarpapers/09-03-12-MP.pdf

Picchio M, Staffolani S (2013) Does apprenticeship improve job opportunities? A regression discontinuity approach. IZA Discussion Paper No 7719. Institute for the Study of Labor, Bonn

Reed D, Liu AY-H, Kleinman R et al (2012) An effectiveness assessment and cost-benefit analysis of registered apprenticeship in 10 states: final report. Mathematica Policy Research, Princeton

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55

Schaeffer CM, Henggeler SW, Ford JD et al (2014) RCT of a promising vocational/employment program for high-risk juvenile offenders. J Subst Abus Treat 46(2):134–143

Skills Development Scotland (2013) Modern apprenticeships outcomes survey 2012. Skills Development Scotland, Glasgow

Skills Development Scotland (2015) Modern apprenticeship employer survey 2015. Report. Skills Development Scotland, Glasgow

Matej Bajgar is an economist at the OECD Productivity and Business Dynamics Division of the Science, Technology and Innovation Directorate. He focuses on using microeconomic data to evaluate impact of public support for training and R&D in the private sector and to analyse firm performance. Before joining OECD as a Young Professional, he was a doctoral student at the University of Oxford, researching performance of manufacturing exporters in the emerging economies. During his doctorate, he worked as a consultant on several projects in developing countries, including a randomised evaluation of a government social programme in Lesotho. He holds degrees from the Charles University in Prague and the University of Oxford.

Chiara Criscuolo is the head of the OECD Productivity and Business Dynamics Division of the Science, Technology and Innovation Directorate. Since joining in 2009, Chiara has also worked on climate change, innovation policies and measurement. She is co-ordinating two large cross-country microdata projects on employment dynamics and on productivity. She recently co-authored a report on the Future of Productivity. Prior to joining OECD, she was Research Fellow at the Centre for Economic Performance, London School of Economics. She has published widely in the field of productivity, innovation and international trade. She holds a doctoral degree in Economics from University College London.

Pay for Performance in Primary Care: The Use of Administrative Data by Health Economists

Rita Santos, Sara Barsanti, and Chiara Seghieri

1 Introduction

Health economists have evaluated pay for performance (P4P) schemes to assess if they are efficient and effective and if they provide positive incentives regarding patient health and inequality reduction.

Researchers and policymakers are concerned with efficiency, effectiveness, value and behaviour in the production and consumption of health and health care. As will be demonstrated in this chapter, P4P schemes and performance measurements of health care systems more generally are usually evaluated on those criteria.

Furthermore, studies on P4P mainly use various administrative datasets that are linkable. The fact that it is possible to link different datasets is essential, since health care, well-being and economics are parts of a multifaceted society.

National statistics institutes produce publicly available socio-economic and demographic data for each country at different geographical scales. This is fundamental to understanding each country as a whole, but also the wide variation within a country and even within a city. It is essential to take these characteristics into account when analysing the performance of primary care providers, because those characteristics not only shape the health needs of their patients but are also a proxy for their responsiveness to continuity of care.

Since P4P is based on performance indicators, health authorities collect information on these for each primary care provider, which is publicly available in most

R. Santos (✉)
Centre for Health Economics, University of York, York, UK
e-mail: rita.santos@york.ac.uk

S. Barsanti · C. Seghieri
Laboratorio Management e Sanità, Institute of Management, Pisa, Italy

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_19

countries. The information relates not only to the performance indicators but also to the population the provider serves and the main characteristics of the provider (e.g. number of doctors).

Once a country or a health region starts a P4P scheme in primary care, it is important to understand the benefits it brings to the health care system in the short and long term. This chapter will describe primary care P4P evaluation and incentives schemes in Italy and England.

The chapter describes the experience of the application of two management tools based on microdata to assist decision makers in monitoring and improving health care system performance, with a particular focus on primary care. In particular, the first case study describes the experience of P4P in general practice in England. The introduction of P4P in England is reported, along with its impact in improving chronic care conditions, measured through indicators of preventable emergency admissions (i.e. ambulatory care-sensitive conditions (ACSCs)), since these emergency admissions are likely to be reduced by improvement in the quality of primary care. The definition of ACSCs is internationally debated, and various definitions and measuring methods have been developed. However, it remains one of the common indicators derived from administrative data to compare primary care quality across providers.

The second case study refers to the development of a performance evaluation system (PES) for general practitioners in Tuscany Region (Italy). This section describes in detail how performance indicators from administrative and ad hoc data sources are measured, shared and made available to practitioners and policymakers to support performance improvement and alignment with the strategic goals of the health care system. Indeed, transparent, systematic benchmarking of performance and feedback mechanisms for health professionals are of strategic significance in designing effective P4P systems, based on reliability, recognition and trust of health care providers.

Considering these two case studies, the reader can be the judge of the importance of exploring and combining different administrative datasets used in the studies. Moreover, the England case study shows some results in terms of counterfactual impact assessment, considering the reduction of some sets of ACSC emergency admissions in England; the Italian case study focuses mainly on how micro-administrative data can support policymakers and public management decision-making, considering the performance reporting system for general practitioners and the first results in terms of improvements in the quality and appropriateness of care.

The plan for the remainder of the chapter is as follows. Section 2 discusses P4P schemes at international level, focusing on performance indicators, incentive schemes and data sources. Section 3 provides evidence on the English P4P scheme and the use of preventable emergency admissions to evaluate the impact of the scheme. Section 4 is dedicated to the Tuscany Region PES at primary care level, as an example of a reporting system for primary care using administrative data sources. Section 5 provides conclusions.

2 Pay for Performance in Primary Care

Within health care systems, countries use a combination of payment methods for doctors, including capitation formulae, salaries, fees for services and P4P schemes, to compensate for weakness associated with single payment approaches and suboptimal health care delivery (Eijkenaar et al. 2013).

Considering the principal–agent theory, when the interests of a principal (i.e. the policymakers) and its agents (i.e. family doctors, also known as general practitioners (GPs)) are not perfectly aligned, P4P may align agents' incentives with the principal's preferences. By making agents' reimbursement depend on their performance with respect to certain performance indicators, the hope is that agents respond by increasing effort on tasks valued by the principal (Prendergast 1999; Anell et al. 2015). Although P4P systems and PESs are used internationally, systematic review of P4P schemes report that the effects of the schemes remain largely uncertain (Houle et al. 2012).

In considering the P4P scheme, there is a need to emphasise that measuring primary care performance is a complex issue, since this level of care encompasses a myriad of activities and its functions differ considerably across countries and also in terms of provider payment and contractual schemes (Roland 2004). Furthermore, different primary care systems also have important effects on data collected and measured.

In P4P, care providers (i.e. family doctors or GPs) receive explicit financial incentives based on their scores on specific measures, which may pertain to certain performance expectations and standards in key domains such as clinical quality, resource use, access to services and patient-reported outcomes. In addition to the United States (USA), where P4P has become widespread, P4P programmes are being implemented in many other countries, including the United Kingdom, Canada, Australia, New Zealand, Taiwan, Israel, France, Italy and Germany, and in both primary care and inpatient care (Milstein and Schreyoegg 2016). Of all P4P physician's programmes in the US 10 years ago, more than 41% were related to primary care; the rest (59%) targeted both primary care physicians and specialists, or specialists only (Eijkenaar et al. 2013).

As reported by Sutherland et al. (2012), 'there is no accepted international definition of pay-for-performance', which may explain the degree of heterogeneity seen among P4P programmes. Moreover, P4P in primary care is context dependent and comparison can be difficult, and the way in which P4P schemes are designed and implemented can affect both the incentives provided and how physicians respond to them (Mehrotra et al. 2010; Eijkenaar et al. 2013).

2.1 *Performance Domains and Indicators*

Primary care P4P programmes usually select measures relating to conditions that are widespread (such as cardiovascular disease) and contribute significantly to the overall burden of disease (such as coverage of vaccinations). In general, clinical indicators refer to compliance in following guidelines for common chronic care conditions, such as diabetes and cardiovascular disease. Considering efficiency, primary care providers may be incentivised for patients requiring below average levels of specialist services, inpatient hospital admissions, drugs prescription in terms of generic drugs consumption (e.g. statins and proton pump inhibitors (PPIs) dispensed) and pharmaceutical expenditure. France and New Zealand have specific targets for reducing pharmaceutical expenditure. Influenza vaccination rates for the elderly (or other target groups, such as children) and cancer screening participation are the main performance measures for preventative care. In the case of nonclinical indicators, performance measurement generally reflects the use of information and communication technology (ICT) (i.e. for registers, appointments and other facilities). Finally, in some countries, such as New Zealand, some indicators are measured separately for high-need sections of the population, and reduction in health inequality and disparities in access to services is also considered a general aim of the P4P scheme (Cashin et al. 2014).

2.2 Data Source

The role of data, information and reporting systems is crucial. Administrative health care data are collected by private health maintenance organisations or governmental institutions, for both managerial and epidemiological reasons (Gini et al. 2013). The use of administrative data to develop performance indicators related to primary care has been increasing over the years: case finding and algorithms are developed to estimate the prevalence of chronic care conditions (CCCs) and the quality of care for the same conditions. The content of databases varies from country to country: they may contain records collected at hospital discharge or during visits to GPs or specialists, or they may relate to drugs prescriptions or diagnostic procedures (Gini et al. 2013). In Canada, Sweden and the USA, administrative databases contain diagnosis codes from inpatient and outpatient care, enabling the estimation of CCC prevalence. In France, where only drugs prescriptions are available, the prevalence of diabetes, for example, is calculated using records for the prescription of anti-diabetic drugs. However, some authors (Green et al. 2012; Gini et al. 2013) are concerned that indicators estimated using administrative databases might not reflect the actual compliance with standard of care for chronic care patients.

In general, as P4P programmes evolve, data sources move from claim data to information directly derived from general practice. In the United Kingdom (UK), New Zealand and other countries, significant investments are being made in infrastructure for data collection (Eijkenaar 2012).

2.3 Incentive Payments

P4P programmes tend to have small rewards as a share of GPs' income; in terms of bonus payment, the percentages are generally 5% or less, with the exceptions of programmes in the UK (about 25%), Turkey (20%) and France (10%). Relatively low payments seem to be preferred because they are aligned with professionals' norms and values (Eijkenaar et al. 2013). As Cashin et al. (2014) emphasise, the incentive is more powerful if it increases as performance improves; almost all P4P in primary care uses higher payment rates for higher achievement levels. There are three main approaches to measuring achievement against performance (Cashin et al. 2014): (1) the absolute level of a measure towards a certain standard; (2) the change in terms of improvement of a measure; and (3) the relative ranking of a measure among the providers.

In general, participation in P4P is on a voluntary base (Eijkenaar 2012). For physicians participating in a P4P scheme, the rate varies from 99% in the UK, with the Quality and Outcomes Framework (QOF), to 80% in Poland.

3 The Case Study: Pay for Performance in Primary Care in England

The National Health Service (NHS) is a tax-financed system and free at point of demand (apart from a small charge for dispensed medicines). NHS primary care is provided by family doctors, known as GPs, who are organised in small surgeries known as general practices. All residents in England are entitled to register with a general practice, and have incentives to do so, as the practices provide primary care and act as the gatekeeper for elective (nonemergency) hospital care.

Most general practices are partnerships owned by GPs and they have, on average, five GPs (four full-time equivalents (FTEs)). They employ other medical staff, including nurses (an average head count (HC) of three, or two FTEs), direct patient care staff (average HC of two, or 1.3 FTEs) and administrative staff (average HC of 12, or eight FTEs) and have around 7500 patients (NHS Digital 2016) The NHS contracts with the practice rather than with the individual GPs. Practices are paid through a combination of lump sum payments, capitation, quality incentive payments and items of service payments. Quality incentives from the P4P scheme, the QOF (Roland 2004), generate a further 15% of practice revenue. Practices are reimbursed for the costs of their premises but have to fund all other expenses, such as hiring nurses and clerical staff, from their revenue.

3.1 *The English Pay for Performance System: The Quality and Outcomes Framework*

The NHS introduced the P4P contract for general practices—the QOF—in 2004/5.[1] This contract was intended to increase GPs' pay by up to 25%, depending on their performance with respect to 146 quality indicators relating to clinical care for ten chronic diseases, organisation of care and patient experience (Roland 2004).

Most of the organisation of care and patient experience indicators rely on a simple characteristic of the practice. For example, organisation of care Record 15 indicator in 2004/5 stated 'The practice has up-to-date clinical summaries in at least 60% of patient records', and patient experience indicator PE2 specifies 'The practice will have undertaken an approved patient survey each year'. While the clinical indicators can take the same form (e.g. Coronary Heart Disease indicator one states that 'The practice can produce a register of patients with coronary heart disease'), most of them are set to vary the attribution of points according to the proportion of patients for whom they achieve each target. On the latest indicators, points are awarded on a descending scale within the payment range. The payment range for each clinical indicator is between the minimum and maximum threshold. A practice whose achievement on a clinical indicator is below 25% does not earn a single point, while one that reaches the maximum threshold achievement (set between 55% and 90% in 2004/5) earns the maximum number of points.

Since some patients might not be eligible for specific indicators, practices are allowed to exclude them given certain circumstances. The NHS publishes the reasons for patient exclusion (in an 'exception report'), along with the P4P indicators.

The P4P indicators and disease groups have changed over the years. In 2004/5, there were four domains: clinical, organisational, patient experience and additional services. The clinical domain had 76 indicators in 11 areas (Coronary Heart Disease, Left Ventricular Dysfunction, Stroke and Transient Ischaemic Attack, Hypertension, Diabetes Mellitus, Chronic Obstructive Pulmonary Disease, Epilepsy, Hypothyroidism, Cancer, Mental Health and Asthma), while the organisational domain had 56 indicators in five areas (Records and Information, Patient Communication, Education and Training, Medicines Management and Clinical and Practice Management), the patient experience had four indicators in two areas (Patient Survey and Consultation Length), and the additional services domain had ten indicators in four areas (Cervical Screening, Child Health Surveillance, Maternity Services and Contraceptive Services). The organisational domain was retired in 2013/14. In 2012/13, the organisational domain still had 42 indicators and the average points per practice was 247.2 out of a maximum of 254 (HSCIC 2013). The patient experience domain was reduced to one indicator in 2013/14 (the average points per practice was 99.7 out of 100) and retired in 2014/15. The additional services domain was renamed

[1] The UK fiscal year starts on 1 April and ends on 31 March of the following year.

‘public health—additional services’ in 2013/2014 and had nine indicators in the same four areas. However, from 2014/15, the public health—additional services domain includes only five indicators in two areas (Contraception (age below 55) and Cervical Screening (age 25–64)), which means that Child Health Surveillance and Maternity Services are no longer incentivised by the QOF. The clinical indicators also underwent changes in terms of new clinical areas. In 2015/16, the clinical domain has 65 indicators in 19 areas (Asthma, Atrial Fibrillation, Cancer, Coronary Heart Disease, Chronic Kidney Disease, Chronic Obstructive Pulmonary Disease, Dementia, Depression, Diabetes Mellitus, Epilepsy, Heart Failure, Hypertension, Learning Disabilities, Mental Health, Osteoporosis, Palliative Care, Peripheral Arterial Disease, Rheumatoid Arthritis, and Stroke and Transient Ischaemic Attack). A new public health domain was added to the QOF in 2013/14, with nine indicators in four areas (Blood Pressure (age above 40), Cardiovascular Disease—Primary Prevention, Obesity (age above 16) and Smoking (age above 15)).

3.2 *The Impact of the Quality and Outcomes Framework on Preventable Emergency Admissions*

The QOF was expected to raise clinical quality and, in particular, to prevent chronic condition events. However, Roland (2004) highlights that in addition to the expected benefits, some negative unintended consequences might also be expected, e.g. a focus on financial reward that is tied to specific tasks and a reduction in the quality of care for conditions not included in the incentive system.

The first studies on the impact of the QOF showed that English general practices had high levels of achievement in the first year of the scheme (Doran et al. 2006), but this did not imply that patients had better continuity of care (Campbell et al. 2009); on the contrary, once the targets were reached, the improvement in some conditions slowed.

Preventable emergency admissions, also known as ACSC emergency admissions, are defined as hospital emergency admissions that could be prevented or reduced through management of the acute episode in the community, or by preventive care (Purdy et al. 2009). Therefore, the ACSCs are a set of disease groups or, more precisely, of diagnosis codes using the tenth revision of the medical classification named International Statistical Classification of Diseases and Related Health Problems (*ICD-10*).

The records are usually retrieved from the patient’s hospital admissions administrative data, which include several ICD-10 codes (including the principal diagnosis for which the patient was admitted), treatment codes, admission and discharge data, the secondary care provider code, the diagnosis-related group, the age and gender of the patient, the geographical area of his/her address and, depending on the country, his/her primary care provider. ACSCs may reflect the suboptimal capacity of health services delivery to effectively prevent, diagnose, treat and/or manage these conditions in primary care settings. ACSC rates are, therefore, inversely correlated with primary care performance, and studies have shown an inverse correlation between treatment guidelines adherence and ACSC inpatient rate (WHO 2016).

To understand the impact over time of the QOF on hospital emergency admissions, Harrison et al. (2014) analysed the time trends, between 2004 and 2011, of ACSC emergency admissions that were incentivised[2] by the QOF and those that were not.

Harrison et al. (2014) report a clear increase in non-incentivised preventable emergency admissions (and non-preventable emergency admissions) and a contrasting decrease in incentivised preventable emergency admissions. Analysis of the period extending from 5 years before the introduction of the QOF scheme to 7 years after its introduction revealed that the emergency admissions rates for all conditions increased by 34% between 1998/99 and 2010/11, while non-incentivised ACSC emergency admissions increased by 39% and non-ACSC emergency admissions rose by 41%. In contrast, incentivised ACSCs decreased by 10%. This decrease is even more important given that the rate of emergency admissions for incentivised ACSCS had been increasing by 1.7% per year before the introduction of the QOF scheme. The fact that the trends in incentivised ACSCs fell by 2.7–8% compared with non-incentivised ACSCs, and by 2.8–10.9% compared with non-ACSCs, between the first year of the QOF (2004/5) and (2010/11), shows that targeting chronic disease groups in primary care can reduce the emergency admission burden on resources and health care costs. The difference between the trends shows that the primary care P4P scheme had a significant impact on hospital emergency admissions.

To assess the impact of general practice characteristics on preventable emergency admissions, Gravelle et al. (2015) linked several administrative datasets between 2006/7 and 2011/12, including datasets on the QOF, ACSCs emergency admissions (from Hospital Episode Statistics), general practice workforce and list characteristics, patient satisfaction and catchment area characteristics. The authors show that the number of GPs and the different practice quality indicators have a significant negative effect on the number of ACSC admissions. The proportion of female GPs has an unexpectedly positive effect on admissions. The effect of GP average age is nonlinear, with the effect of average age declining and then reversing around 55 years of age. Neither the proportion of non-UK qualified GPs nor the proportion of salaried GPs has any significant effect. The patient-reported measures of ability to obtain urgent or advance appointments have negative coefficients, suggesting that this type of access reduces ACSC admissions. Being able to see a preferred GP also reduces admissions, hinting at the beneficial effects of continuity or good

[2] Harrison et al. (2014) included in the incentivised ACSCs group disease groups that had clearly been continuously incentivised under the QOF since the introduction of the scheme in 2004, and, as non-incentivised ACSCs, the remaining disease groups that were not targeted under the QOF at any time between 2004/5 and 2010/11.

interpersonal relations between patients and GPs. The more objective measure of clinical quality derived from the QOF also has a small negative effect on ACSCs. A practice that has a better QOF clinical quality achievement has significantly fewer ACSC emergency admissions, but the impact is small.

Given the opportunity to link administrative datasets on patient hospital admissions to general practice information, and general practice to census information, Kasteridis et al. (2016) and Goddard et al. (2016) examined the impact of the introduction of the 2006 QOF indicator for dementia on discharge destination and length of stay for patients admitted for dementia and on ACSC emergency admissions from 2006/7 to 2010/11, respectively. More precisely, Kasteridis et al. (2016) analysed if patients registered with GP practices that have better QOF indicator scores for the annual dementia review have a smaller likelihood of a care home placement following an acute hospital emergency admission. The major predisposing factors for institutionalisation in a care home were older age, female gender and the need factors of incontinence, fall, hip fracture, cerebrovascular disease, senility and total number of additional comorbidities. Over and above those factors, the dementia QOF review had no significant impact on the likelihood of care home placement for patients whose emergency admission primary diagnosis was dementia, but there was a small negative effect if the emergency admission was for an ACSC, with an odds ratio of 0.998. On the other hand, Goddard et al. (2016) found a significant and negative effect of the Dementia QOF indicator on length of stay among urgent admissions for dementia. Patients discharged to the community had significantly shorter hospital stays if they were cared for by practices that reviewed a higher percentage of their patients with dementia. However, this effect is not significant for patients discharged to care homes or who died in hospital. The authors also report that longer length of stay is associated with a range of comorbidities, markers of low availability of social care and intensive provision of informal care. Dusheiko et al. (2011) investigated another link between the QOF and a specific ACSC. The authors explored the association between general practices' quality of diabetic management, given by QOF indicators, and emergency admissions for short-term complications of diabetes between 2001/2 and 2006/07, i.e. before and after the introduction of the QOF. They reported that practices with better quality of diabetes care had fewer emergency admissions for short-term complications of diabetes. However, they did not find an association with hypoglycaemic admissions.

Some studies have also used cross-sectional analysis to assess the impact of the QOF in specific ACSCs. For example, Calderón-Larrañaga et al. (2011) analysed the association between the specific QOF Chronic Obstructive Pulmonary Disease (COPD) indicators and COPD hospital admissions in 2008/9 (the year before the influenza pandemic). The authors reported that smoking prevalence and deprivation were risk factors for admission, while the QOF indicator for patients with COPD who had received an influenza immunisation (an increase in the QOF indicator of 1% would be expected to decrease COPD admissions by a factor of 0.825), patient satisfaction with ability to book a GP appointment within 2 days and the number of GPs per 1000 patients in the practice were protective factors for COPD admissions.

The early impact of the QOF on angina and myocardial infarction 2006/7 hospital admissions was analysed by Purdy et al. (2011). While a higher overall clinical QOF score was associated with lower rates of admissions for angina and myocardial infarction, the four specific coronary heart disease indicators were not.

Overall, the NHS England primary care P4P scheme—the QOF—had a positive impact on the reduction of ACSC emergency admissions, overall and for specific conditions. The reductions were clear for the conditions incentivised by the QOF, especially during the period soon after the introduction of the scheme.

4 The Case Study: Pay for Performance in Primary Care in Italy

4.1 Health Care System and GPs in Italy

The Italian public health care system is inspired by the Beveridge model, and it is characterised by public taxation funding, free access at the point of delivery (with some copayments for specific services), and political control over providers.

In the Italian national health system, GPs are the first contact for most common health problems and act as gatekeepers for drug prescription and for access to secondary and hospital care. Their activities and responsibilities have three levels of governance (Barsanti et al. 2014): (1) the national level (through the 'National Agreement' between the central government and the national GPs' trade unions (TUs)); (2) the regional level (through the 'Regional Agreement' between the regional government and the regional TUs); and (3) the local level (through the 'Local Health Authority Agreement', between the local health authority (LHA) managers and the local TUs). Primary care physicians are paid through a combination of methods, and regional and local health authorities have some degree of autonomy in defining additional payment. Each region may introduce economic incentives to complement the national current payment structure. These economic incentives can relate to performance, appropriateness of care or the adoption of patient referral.

The recent health planning legislation (Balduzzi Law No. 189/2012 and the *Patto per la Salute* ('Agreement for Health') 2014–2016) introduces strategies for the organisation of primary care according to operational forms that include a single unit of professional organisation, the *Aggregazioni Funzionali Territoriali* ('Territorial Functional Aggregations') (AFTs). AFTs represent the highest level of general practice organisation, and each serves a population of 30,000 patients, assisted by 25 GPs. Each AFT has a coordinator elected by the GPs. AFTs are expected to apply the philosophy of 'clinical governance' (Scally 1998), whereby GPs have responsibility for continuously improving the quality of their services and safeguarding high standards of care.

Tuscany was the first Italian region to adhere to the new collaborative AFT model. The Tuscany Region has about 4 million inhabitants, and in 2015 its health care system comprised 12 LHAs (merged into three LHAs in 2016) and four teaching hospitals. In 2012, 115 AFTs were created through the Tuscany Regional Agreement and subsequent local agreements (Barsanti et al. 2016). On average, each AFT has 28,000 inhabitants with an average age of 52 years. In Tuscany, there are about 2700 GPs with on average 1100 patients each. The average age of GPs is about 60 years, with men in the majority.

4.2 The Making of the Tuscan Performance Evaluation System (PES) for Primary Care

In 2004, Tuscany Region commissioned the Scuola Superiore Sant'Anna of Pisa to design and implement a multidimensional PES to monitor the results of LHAs in terms of clinical quality and appropriateness, both for the hospital setting and for the district setting (Nuti et al. 2012, 2016). Starting from 2007, the performance indicators within the PES were presented in terms of a benchmark conducted across the health care providers and made available on a web platform for managers and professionals. The PES has now become the core of the P4P scheme of the CEOs of the LHAs. Starting in 2013, selected performance indicators within the PES were also calculated at AFT level to monitor and compare GPs' performance with respect to primary care activities and responsibilities, including (1) management of chronic disease; (2) prevention of avoidable hospital admission and inappropriate diagnostic tests; (3) preventive care and home care for the elderly; (4) drug prescriptions; (5) practice organisation; and (6) patient experience (Barsanti and Nuti 2016).

The Tuscan PES encompasses a large set of indicators grouped into about 25 indexes and classified in five dimensions (Table 1). Indicators are defined in regular meetings between the regional administration and representatives of general practices, including the perspectives of both managers and clinicians. The main source of data for clinical indicators are local health administrative data, which are centrally collected by the regional administration. Clinical indicators are measured using regional administrative data, which comprise electronic records of all inpatient and outpatient activity, as well of pharmaceutical consumption among all residents from 2009 to 2016. Patient experience and GP organisation measures at AFT level are collected through sample surveys (De Rosis and Barsanti 2016). Research and evaluators utilise various sources of microdata through data record linkage, including:

1. Linkage between patients, their usual GP and the GP's AFT.
2. Data linkage at individual (patient) level across different administrative databases (i.e. inpatient, outpatient and drug consumption data), to measure performance indicators along the different care pathways for chronic patients (e.g. process indicators for the care of diabetic patients).

Table 1 Domains, indicators and data sources of the PES for GPs in Tuscany

Domains	Example of indicators	Data source
Regional primary care strategy compliance: to guarantee that strategic regional goals are pursued in the appropriate time and manner, focusing on primary care setting	– Influenza vaccination rate for older people – Percentage of older people with home care assistance	Ad hoc data sourced from GPs and LHAs, and administrative health data flow
Management of chronic care patients and preventable inpatients: focusing on quality, appropriateness and continuity of care for patients with chronic care disease	– Ambulatory care-sensitive condition inpatient rate (standardised by sex and age) – Access to emergency department for minor diseases (standardised by sex and age) – Percentage of patients with heart failure receiving angiotensin-converting enzyme (ACE) inhibitor therapy – Percentage of patients with heart failure receiving beta blocker therapy – Percentage of patients with diabetes receiving haemoglobin testing – Percentage of patients with diabetes receiving eye checks by an oculist	Administrative health data flow
Patient satisfaction: patient experience and level of satisfaction with GPs	– Percentage of patients satisfied with their GP in terms of communication, general assistance, time for visiting, collaboration	Statistical sample survey
GPs satisfaction: results of surveys on the satisfaction level and experience of GPs with their general practice organisation	– Meeting to share clinical guidelines – Frequency of clinical audit	Survey of all GPs
Diagnostic and pharmaceutical care: focusing on appropriateness and compliance of care in drugs prescription and diagnostic visits	– Nuclear Magnetic Resonance (NMR) rate for older people – Consumption of specific drugs (statins, PPI) – Drug prescriptions for generic drugs	Administrative health data flow

Note: *GP* General Practitioner, *LHA* Local Health Authority, *ACE* Angiotensin-Converting Enzyme, *NMR* Nuclear Magnetic Resonance, *PPI* Proton-pump inhibitors

Indicators were selected on the basis of the following criteria (Vainieri et al. 2016): (1) scientific soundness of the indicators in terms of their validity and reliability; (2) feasibility of obtaining regionally comparable data for all 115 AFTs; (3) ability to provide a comprehensive overview of primary care.

Each indicator considers regional, national or literature-based standards in order to assess performance. Where standards are absent or difficult to set, the regional variation of the practice is taken into account, benchmarking performances with the quintile distribution. Moreover, each performance indicator is measured at a different level of governance: regional level, LHA level, AFT level and individual GP level.

PES indicators that are considered as evaluation measures are assigned performance assessment ratings for benchmarking reporting across AFTs. For each evaluation measure, five performance levels are derived for defining the performance of each AFT, from worst to best. These five evaluation tiers are associated with different colours, from dark green (excellent performance) to red (poor performance).

The PES considers three main perspectives of performance evaluation for each indicator:

- Performance assessment with respect to a standard derived either from the literature or from agreements at local and/or regional level. Data are displayed as histograms reporting the performance value of all 115 AFTs.
- Variability among AFTS and LHAs (Fig. 1) represented by a box plot with value of AFTs grouped by LHAs and by a boxplot with individual GPs grouped by AFT (Fig. 2).

Considering all the performance indicators, the summarising reporting system is visually represented by a 'target' diagram, which is divided into five coloured evaluation bands. Every year each AFT receives its own target and the more the AFT is able to reach the yearly defined objectives, the closer the performance indicators are to the centre (the dark-green area). Scarce and critical performance results are represented by indicators, which are positioned far from the centre, in the red area (Fig. 3).

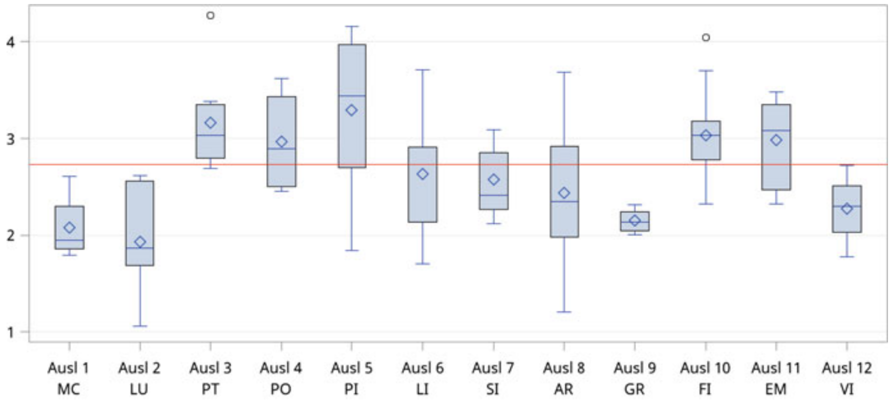


Fig. 1 Box plot of ACSC hospital admission between local health authorities in 2015. Source: Barsanti and Nuti (2016)

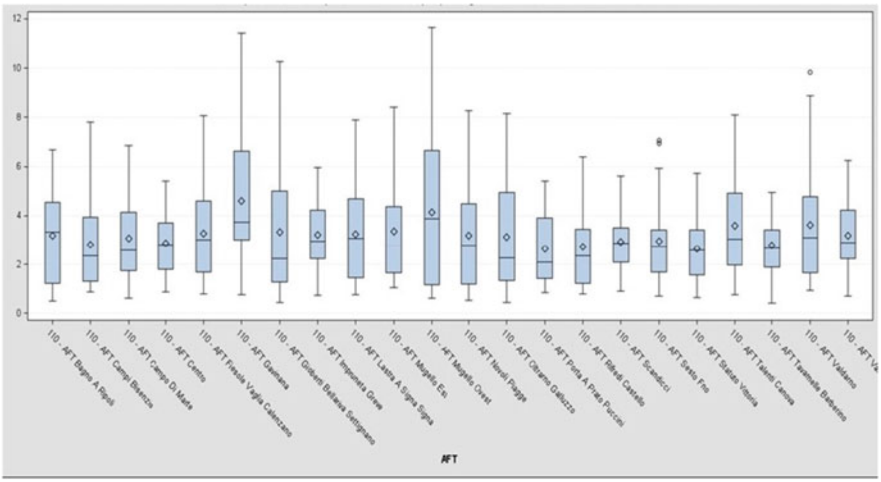


Fig. 2 Box plot of ACSC hospital admissions between AFTs of local health authorities (LHAs) in Florence in 2015 (in 2015 there were 22 AFTs grouped in the LHA of Florence). Source: Barsanti and Nuti (2016)

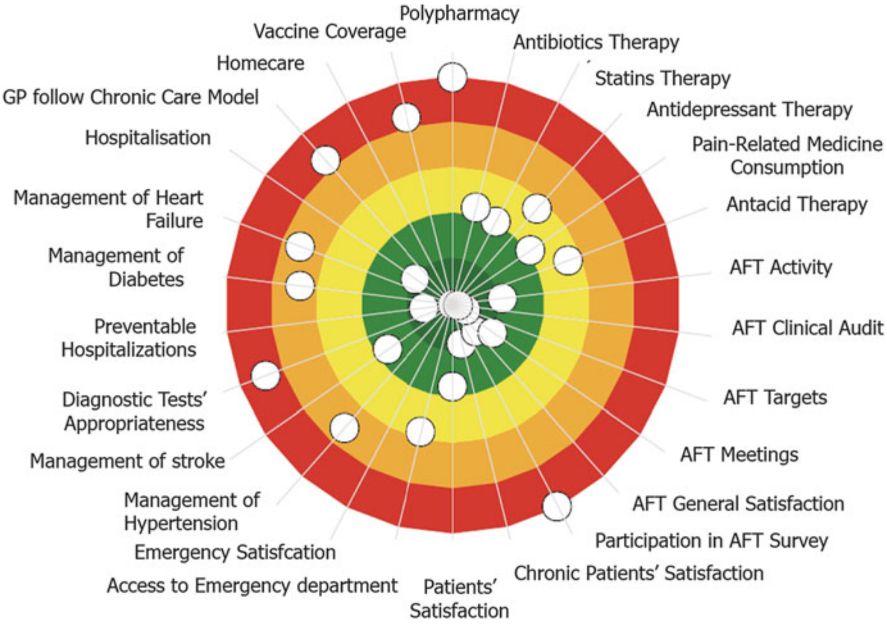


Fig. 3 Example of the system of reporting performance indicators AT AFT LEVEL in 2014'. Source: Barsanti and Nuti (2016)

4.3 The PES and the P4P Programme in Tuscany Region: First Results

A recent paper (Nuti et al. 2016) examines the various governance models in the health care sector that the Italian regions have adopted, and investigates the PESs associated with them, focusing on the experience of a network of ten regional governments that use the same health care PES as Tuscany Region. Considering 14 indicators measured in 2007 and in 2012 for all the regions, the study shows how different performance evaluation models are associated with differences in health care performance and whether or not the use of a PES has made any difference to the results achieved by the regions involved. In particular, Tuscany Region registered a high performance in 2007 and was still offering good general assistance in 2012, even improving both hospital and primary care processes (Nuti et al. 2016). In this sense, the authors conclude: systematic benchmarking and public disclosure of data based on the Tuscany PES are powerful tools that can guarantee sustained improvement of health care systems, if they are integrated with the regional governance mechanisms.

With regard to primary care, the P4P scheme for general practice in Tuscany Region has recently been developed following the P4P scheme for the chief executive officers of LHAs (Nuti et al. 2012). At this stage in the reform of primary care (see the constitution of the AFT), policymakers and professionals have some degree of autonomy in the design of the P4P programme, in terms of selection of performance indicators and incentives (Barsanti et al. 2014). Therefore, different groups of GPs might be incentivised on different sets of PES indicators. An analysis of the 12 Local Health Agreements for Tuscany Region in 2015 shows that almost 50% of the PES indicators were used also in the local P4P programmes for primary care. In particular, all indicators measuring pharmaceutical care, preventable hospital admission (ACSC inpatient rate) and management of chronic care disease are used to incentivise GPs. Each AFT has its own standard to achieve, based on the performance of the previous year. Usually, incentives are set both at AFT and at individual GP level.

Although the use of PESs in primary care and the linkage to GP incentives have only recently been introduced, and although the data do not allow for measuring any causal effect of P4P on performance, preliminary results show a positive impact of P4P on the quality and appropriateness of primary care in Tuscany (Barsanti et al. 2014), as measured by significant improvement over the years of performance of selected indicators included in the P4P system. About 50% of indicators improved from 2014 to 2015 at regional level (Barsanti and Nuti 2016).

In this sense, three selected primary care performance indicators used in the Tuscany GP PES are compared, in order to assess improvements over the years 2013–2015, considering:

1. The percentage of elderly people who received home care.
2. The rate of hospital admissions for ACSC per 1000 inhabitants (standardised by age and sex).

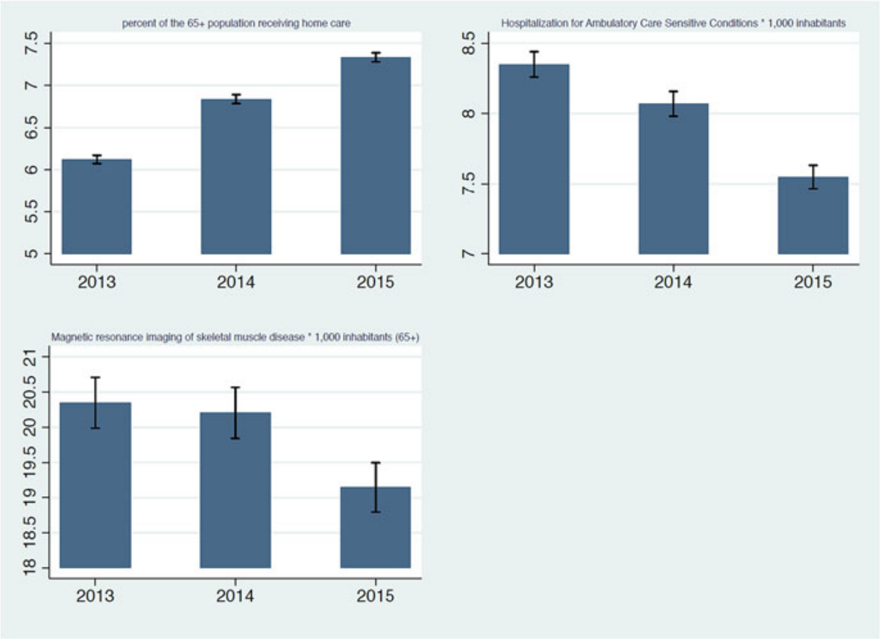


Fig. 4 Percentage receiving home care, ACSC inpatient rate standardised by age and sex and magnetic resonance imaging of skeletal muscle among the population aged over 65 years: trends 2013–2015 and confidence intervals

3. The rate of magnetic resonance imaging for musculoskeletal disease per 1000 elderly people (standardised by age and sex).

All the selected indicators show significant improvements, considering trend and confidence intervals (Fig. 4).

In Tuscany, ACSC inpatient rates decreased by 10% between 2013 and 2015 (from 8.35 per 1000 residents in 2013 to 7.54 per 1000 residents in 2015), whereas, in contrast, the number of elderly people (aged 65+) receiving home care increased by 20% (from 6.12% in 2013 to 7.33% in 2015). With regard to indicators of appropriateness, although the rate of magnetic resonance imaging for musculoskeletal disease among elderly people (those aged 65+) did not change appreciably from 2013 to 2014, it decreased by 6% from 2013 to 2015 (from 19.14 per 1000 elderly people in 2013 to 20.35 per 1000 in 2015) (Fig. 4).

Finally, considering the use of the PES, GPs use the PES results during clinical audits to formulate evidence-based improvement strategies for the care of their patients.

The Tuscan PES for primary care is playing an increasing role, providing valuable information to GPs and local decision-makers to support quality improvements, to define priorities and to set appropriate targets.

5 Conclusion

As we enter the ‘big data’ era, it is important to rethink the usage of administrative data to allow for timely evidence-informed clinical and policy decision-making. Although P4P has high face validity, the evidence for the effectiveness of such schemes in improving quality remains mixed (Rosenthal and Frank 2006; Mehrotra et al. 2010; Eijkenaar et al. 2013). A recent literature review on the effects of P4P programmes found 58 studies related to primary care P4P scheme and results (Mendelson et al. 2017). The authors found low-strength evidence that P4P programmes may improve process-of-care outcomes over the short term (2–3 years). Many of the studies reporting positive findings were conducted in the United Kingdom, where incentives are much larger than any P4P programmes in the United States. The largest improvements were seen in areas where baseline performance was poor.

This chapter describes two experiences of using performance measurement and P4P, whose efficacy and impact can be analysed only by linking data from regional population and administrative databases (i.e. from the P4P to the population census). The P4P schemes had a positive impact on the quality of primary care in both countries. However, more research is needed to understand the expected value of an indicator in terms of its impact on quality and its lifespan.

In England, the primary care P4P scheme—the QOF—had a positive impact on quality. However, practices achieved, on average, 958.7 points, representing 91.3% of the total 1050 points available, in the first year of the scheme. After changes to the P4P domains, clinical areas and indicators over the following 11 years, practices still had a high achievement, with an average of 532.9 points out of 559 in the clinical domain in 2015/16. In addition to the achievement of the P4P indicators, there was also a reduction in ACSC emergency admissions, especially for incentivised disease groups. Whether or not the long-term effects will compensate for the costs of P4P is a topic for further research.

The PES of Tuscany Region is presented as an example of how the combination of performance measurement and explicit incentives can be effectively used to promote accountability and to improve the quality of care in a regional health system. Moreover, the mix of systematic benchmarking between primary care providers over the years and the identification of best practices offer an overall strategic framework of evidence-based information to be used by professionals in everyday practice, and by decision-makers to define priorities and set performance targets.

It is essential to understand the contribution of the measurement of performance and use of incentives in primary care to the improvement of health care quality, to the reduction of unwarranted variation and to the ultimate decrease of secondary care burden. General practitioners and family practices are on the front line of prevention and treatment, and are key in any national health system to achieve its goal: a healthy population.

Acknowledgements Rita Santos was funded by a National Institute for Health Research Doctoral Fellowship (DRF 2014-07-055). The PES of the Tuscan health care system was founded by Tuscany Region through an agreement with Scuola Superiore Sant'Anna di Pisa. The views expressed are those of the authors and not necessarily those of the funders.

References

Anell AD, Jens D, Ellegård LM (2015) Can pay-for-performance to primary care providers stimulate appropriate use of antibiotics? School of Economics and Management Working Paper 2015:36, Lund University, Lund

Barsanti S, Nuti S (2016) Il sistema di valutazione della performance delle AFT toscane. Edizioni del Gallo, Perugia

Barsanti S, Bonciani M, Pirisi L et al (2014) Trade union or trait d'union? Setting targets for general practitioners: a regional case study. Paper presented at the International Research Society for Public Management Conference 2015, University of Birmingham, 30 March–1 April 2015. http://irspm2014.com/index.php/irspm/IRSPM2015/paper/viewFile/1460/585

Barsanti S, Bonciani M, Vola F (2016) Innovatori, indecisi, bisognosi o autonomi: I medici di medicina generale tra integrazione e accountability 2016. Mecosan 98:9–39. https://doi.org/10.3280/MESA2016-098002

Calderón-Larrañaga A, Carney L, Soljak M et al (2011) Association of population and primary healthcare factors with hospital admission rates for chronic obstructive pulmonary disease in England: national cross-sectional study. Thorax 66(3):191–196

Campbell SM, Reeves D, Kontopantelis E (2009) Effects of pay for performance on the quality of primary care in England. N Eng J Med 361(4):368–378. https://doi.org/10.1056/NEJMsa0807651

Cashin C, Chi Y-L, Smith PC et al (2014) Paying for performance in health care: implications for health system performance and accountability. Open University Press and McGraw-Hill Education, Maidenhead

De Rosis S, Barsanti S (2016) Patient satisfaction, e-health and the evolution of the patient–general practitioner relationship: evidence from an Italian survey. Health Policy 120(11):1279–1292

Doran T, Fullwood C, Gravelle H et al (2006) Pay-for-performance programs in family practices in the United Kingdom. N Engl J Med 355(4):375–384. https://doi.org/10.1056/NEJMsa055505

Dusheiko M, Doran T, Gravelle H et al (2011) Does higher quality of diabetes management in family practice reduce unplanned hospital admissions? Health Serv Res 46(1p1):27–46. https://doi.org/10.1111/j.1475-6773.2010.01184.x

Eijkenaar F (2012) Pay for performance in health care: an international overview of initiatives. Med Care Res Rev 69(3):251–276

Eijkenaar F, Emmert M, Scheppach M et al (2013) Effects of pay for performance in health care: a systematic review of systematic reviews. Health Policy 110(2–3):115–130. https://doi.org/10.1016/j.healthpol.2013.01.008

Gini R, Francesconi P, Mazzaglia G et al (2013) Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey. BMC Public Health 13:15. https://doi.org/10.1186/1471-2458-13-15

Goddard MK, Kasteridis P, Jacobs R (2016) Bridging the gap: the impact of quality of primary care on duration of hospital stay for people with dementia. J Integr Care 24(1):15–25

Gravelle HSE, Martin S, Santos R (2015) GPs, practice quality and emergency hospital admissions for ambulatory care sensitive conditions. Paper presented at the Health Economists' Study Group Meeting, Lancaster University, 7–9 January 2015

Green ME, Hogg W, Savage C et al (2012) Assessing methods for measurement of clinical outcomes and quality of care in primary care practices. BMC Health Serv Res 12:214. https://doi.org/10.1186/1472-6963-12-214

Harrison MJ, Dusheiko M, Sutton M et al (2014) Effect of a national primary care pay for performance scheme on emergency hospital admissions for ambulatory care sensitive conditions: controlled longitudinal study. BMJ 349:g6423. https://doi.org/10.1136/bmj.g6423

Houle SD, McAlister FA, Jackevicius CA et al (2012) Does performance-based remuneration for individual health care practitioners affect patient care? A systematic review. Ann Intern Med 157(12):889–899. https://doi.org/10.7326/0003-4819-157-12-201212180-00009

HSCIC (2013) Quality and outcomes framework achievement, prevalence and exceptions data, 2012/13. Health and Social Care Information Centre, Leeds. http://content.digital.nhs.uk/catalogue/PUB12262/qual-outc-fram-12-13-rep.pdf

Kasteridis P, Mason A, Goddard M (2016) Risk of care home placement following acute hospital admission: effects of a pay-for-performance scheme for dementia. PLoS One 11(5):e0155850. https://doi.org/10.1371/journal.pone.0155850

Mehrotra A, Sorbero ME, Damberg CL (2010) Using the lessons of behavioral economics to design more effective pay-for-performance programs. Am J Manag Care 16(7):497–503

Mendelson A, Kondo K, Damberg C et al (2017) The effects of pay-for-performance programs on health, health care use, and processes of care: a systematic review. Ann Intern Med 166(5):341–353. https://doi.org/10.7326/M16-1881

Milstein R, Schreyoegg J (2016) Pay for performance in the inpatient sector: a review of 34 P4P programs in 14 OECD countries. Health Policy 120(10):1125–1140. https://doi.org/10.1016/j.healthpol.2016.08.009

NHS Digital (2016) General and personal medical services, England, September 2015–March 2016, Provisional experimental statistics

Nuti S, Seghieri C, Vainieri M (2012) Assessment and improvement of the Italian healthcare system: first evidence from a pilot national performance evaluation system. J Healthc Manag 57(3):182–199

Nuti S, Vola F, Bonini A et al (2016) Making governance work in the health care sector: evidence from a 'natural experiment' in Italy. Health Econ Policy Law 11(1):17–38. https://doi.org/10.1017/S1744133115000067

Prendergast C (1999) The provision of incentives in firms. J Econ Lit 37(1):7–63

Purdy S, Griffin T, Salisbury C et al (2009) Ambulatory care sensitive conditions: terminology and disease coding need to be more specific to aid policy makers and clinicians. Public Health 123(2):169–173. https://doi.org/10.1016/j.puhe.2008.11.001

Purdy S, Griffin T, Salisbury C et al (2011) Emergency admissions for coronary heart disease: a cross-sectional study of general practice, population and hospital factors in England. Public Health 125(1):46–54. https://doi.org/10.1016/j.puhe.2010.07.006

Roland M (2004) Linking physicians' pay to the quality of care: a major experiment in the United Kingdom. N Engl J Med 351(14):1448–1454. https://doi.org/10.1056/NEJMhpr041294

Rosenthal MB, Frank RG (2006) What is the empirical basis for paying for quality in health care? Med Care Res Rev 63(2):135–157

Scally G (1998) Clinical governance and the drive for quality improvement in the new NHS in England. BMJ 317:61

Sutherland JM, Repin N, Crump RT (2012) Reviewing the potential roles of financial incentives for funding healthcare in Canada. Canadian Foundation for Healthcare Improvement, Ottowa. http://www.cfhi-fcass.ca/Libraries/Reports/Reviewing-Financial-Incentives-Sutherland-E.sflb.ashx. Accessed 9 Feb 2015

Vainieri M, Vola F, Soriano Gomez G (2016) How to set challenging goals and conduct fair evaluation in regional public health systems: insights from Valencia and Tuscany regions 2016. Health Policy 120(11):1270–1278

WHO (2016) Assessing health services delivery performance with hospitalization for ambulatory care sensitive conditions. World Health Organization Regional Office for Europe, Copenhagen. http://www.euro.who.int/__data/assets/pdf_file/0010/305875/Assessing-HSD-performance-with-ACSH.pdf?ua=1

Rita Santos is a research fellow at Centre for Health Economics of the University of York. She holds an NIHR doctoral research fellowship on “Measuring and explaining primary care quality variation”. In the past years, she worked on GP, gender pay, differences, GP practice choice, hospital competition and GP practices competition. Her main interests are developing new applications of geographical information system for health economics and applying spatial econometric methods to economic theory.

Sara Barsanti is assistant professor of Management at the Laboratorio Management e Sanità of Scuola Superiore Sant’Anna of Pisa (Italy). She has a PhD in Health Care Management. In the past years, she has worked on a variety of health research projects and collaborated with several international institutions. Her research interests focus on organizational model of primary care and performance evaluation system in health care.

Chiara Seghieri is assistant professor of Social Statistics at the Laboratorio Management e Sanità of Scuola Superiore Sant’Anna of Pisa (Italy). In the past years, she has worked on a variety of health research projects and collaborated with several international institution and academic/research centers. Her current research interests include the development and use of statistical methods for measuring and comparing care quality and equity using observational data from both sample surveys and administrative records.

The Potential of Administrative Microdata for Better Policy-Making in Europe

Sven Langedijk, Ian Vollbracht, and Paolo Paruolo

1 Introduction

This chapter looks at the future of evidence-based policy-making in Europe. It takes a bird's-eye view and it uses a stylised approach in tackling the question 'What might policy-making and policy evaluation look like in a hypothetical world of perfect availability of administrative microdata?' It reviews possible answers to this question, as well as related benefits and pitfalls.

Evidence-based policy has, since the 1940s and 1950s, been associated with the rise of empirical social sciences such as sociology, economics, political science and social psychology (Head 2015). By the 1970s, leading social scientists increasingly advocated the importance of rigorous behavioural and experimental methods. Analysis of quantitative data from social experiments was advocated, as was the application of advanced analytical methods to data collected by passive observation; these methods later evolved into the set of tools currently known as quasi-experimental methods. By the beginning of the new millennium, the movement had become known as the 'evidence-based policy' movement (Panhans and Singleton 2017).

While this trend was heterogeneous across regions of the world, in recent decades many government-spending programmes have increasingly been evaluated using quantitative methods. The objective is to better determine what works, and to enhance the efficiency and effectiveness of the programmes in question. These evaluations have often been based on programme data and specifically targeted surveys.

S. Langedijk (✉) · I. Vollbracht · P. Paruolo
Joint Research Centre, Ispra, VA, Italy
e-mail: sven.langedijk@ec.europa.eu; ian.vollbracht@ec.europa.eu; paolo.paruolo@ec.europa.eu

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_20

In a typical experimental evaluation design, a baseline survey is run before the policy intervention, and a follow-up survey is repeated to measure the outcomes for the treated and the control groups. The usual limitations of attrition bias, enumerator bias, reporting bias,[1] etc. apply to survey data; while a few of these sources of bias may remain when using administrative data (see Feeney et al. 2015), recourse to such data can greatly reduce them.

In the European Union (EU), the European Commission has embraced the idea of evaluation for spending programmes in the first decade of the new millennium (Stern 2009). The EU began by setting up evaluation requirements for centrally managed spending programmes, and joint evaluation schemes with authorities in the member states for spending programmes managed and executed at national or regional levels. Examples of the latter are the European Social Fund and the European Regional Development Fund.

The Commission has gradually expanded the scope and coverage of its policy evaluations to regulatory and legislative policies. The Juncker Commission made better regulation one of its core goals, and in 2015 published new guidelines for ex post policy evaluation, with the aim of improving policies, reducing administrative burdens, strengthening accountable government and supporting strategic decision-making (Mastenbroek et al. 2015).

Policy evaluations at first had limited quantification, relying mostly on data aggregated at country level, sometimes combined with targeted ex post surveys of beneficiaries and stakeholders. This often limited the ability to make causal evaluation of what worked and what did not.

Although evaluation techniques and the potential comprehensiveness of data continue to improve, even today, data (un)availability remains the key limiting factor preventing the widespread use of more rigorous approaches. At the same time, public authorities are sitting on a treasure trove of administrative data collected and used for other purposes, such as social security, taxation, education, communal administration, cadastres and company registration.

If these data could be effectively reused, then econometric and statistical techniques would allow disaggregation of the analysis along many dimensions, such as geographical, social and firm level. For socio-economic policies, enhanced data availability could allow policy evaluation centred on the life course of citizens (Gluckman 2017). Detailed and accurate evidence on ‘what works’ would allow a step change in the quality of public services and legislation.

The rest of the chapter is organised as follows. Section 2 considers current trends in evaluation, data gathering and storage, automation of data collection processes, and increased processing and storage capacity. At the end of the section, a number of simplifying assumptions on the extrapolation of current trends are made. Section

[1] There may be incentives for either the subject or the collector of data (or both) to misreport data. For example, an individual may be incentivised to underreport income in an application for social welfare services.

3 illustrates the stylised aspects of a world with perfectly functioning access to microdata for policy research. This abstraction is both useful and valid in relation to making relevant choices for future developments; it can help to get a conceptual grip on the direction in which current trends allow policy-makers to steer evidence-based policy-making. Section 4 looks at the potential pitfalls, while Sect. 5 presents concluding remarks.

2 Trends in Data and Policy and Programme Evaluation

This section describes current trends in microdata availability that may steer future developments in evaluation. The last part of the section presents a set of simplifying assumptions on the continuation of these trends; these assumptions are maintained in the rest of the chapter.

2.1 Increased Availability of (Micro)data

The world is producing increasing amounts of data. A key indicator of increased data production and use by private individuals is, for instance, the trend in global IP (internet) traffic. The industry predicts that this will increase threefold between 2016 and 2021, with the number of devices connected to IP networks at three times the global population by 2021. The Internet of Everything phenomenon, in which people, processes, data and things connect to the Internet and each other, is predicted to show impressive growth; globally, machine-to-machine connections are expected to increase by 140% over the period 2016–2021 (Cisco 2017).

There have been large strides forward in the past two decades with regard to computing power, data storage capacity, analytical techniques and algorithm development. These trends, together with a massive increase in the use of devices connected to the Internet by private citizens, have allowed big tech companies such as Amazon and Google to expand at a dramatic rate. While there is little doubt about the benefits that these innovations have delivered in terms of choice, speed and access to information, citizens' concerns about data privacy and security have, in parallel, become much more visible issues in public policy discourse.[2]

In the governmental sphere, the use (for evaluation and policy research in particular) and the potential interlinkage of administrative data held by public institutions have moved forward at a much slower rate.[3] This may reflect some combination of inertia and data security and privacy concerns. However, some steps forward have been achieved, as illustrated below.

[2] The recent EU General Data Protection Regulation, see https://www.eugdpr.org, is addressing the issue of protection of personal information.

[3] Discussion of the work of signal intelligence agencies is beyond the scope of this chapter.

2.2 Administrative Data Linkage Centres

Over the past decade, centres for linking and granting access to microdata to researchers have been set up in a number of countries. The Jameel Poverty Action Lab (J-PAL) North America, based at the Massachusetts Institute of Technology, has compiled a catalogue of administrative datasets available in the United States.[4] In Europe, most countries provide (limited) access to microdata, with some states providing linking services. Statistics Netherlands,[5] for instance, provides linking services to researchers in the Netherlands and other EU countries.

In the United Kingdom, the government's Economic and Social Research Council funded the Administrative Data Research Network (ADRN), with an initial funding period spanning 2013–2018.[6] Other examples of data access and linkage centres are at Germany's *Institut für Arbeitsmarkt-und Berufsforschung* (Institute for Employment Research) (IAB) (see Chap. 7) and New Zealand's Integrated Data Infrastructure, see Gendall et al. (2018).

2.3 Trends in Economic Publications

The current increased availability of microdata is moving the focus of economic research, which has become increasingly centred on data analysis (Hamermesh 2013). A recent analysis of fields and styles in publication in economics by Angrist et al. (2017) documents a shift in publications and citations towards empirical work, where the empirical citation share is now at around 50%, when compared with the two alternative categories of 'economic theory' and 'econometrics'.

Angrist and Pischke (2017) call for a parallel revision of undergraduate teaching in econometrics, which they argue should be more focused on causal questions and empirical examples, should have a controlled-experiment statistical framework as reference and an emphasis on quasi-experimental tools.

Panhans and Singleton (2017) document the rise of empirical papers and of quasi-experimental methods in economics. They track the citations of quasi-experimental methods in the major economic journals. A similar interrogation of Google Scholar for the words 'counterfactual' or 'counterfactual impact evaluation'

[4] See https://www.povertyactionlab.org/admindatacatalog. All links and Internet searches in this chapter were accessed or processed in November 2017.

[5] For microdata at Statistics Netherlands, see https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research.

[6] The ADRN (https://adrn.ac.uk/) consists of: a coordinating body (the Administrative Data Service); four Administrative Data Research Centres, one in each country in the United Kingdom; national statistics authorities, government departments and agencies (the data providers), the Economic and Social Research Council (the funders), and the UK Statistics Authority (which leads the network's board).

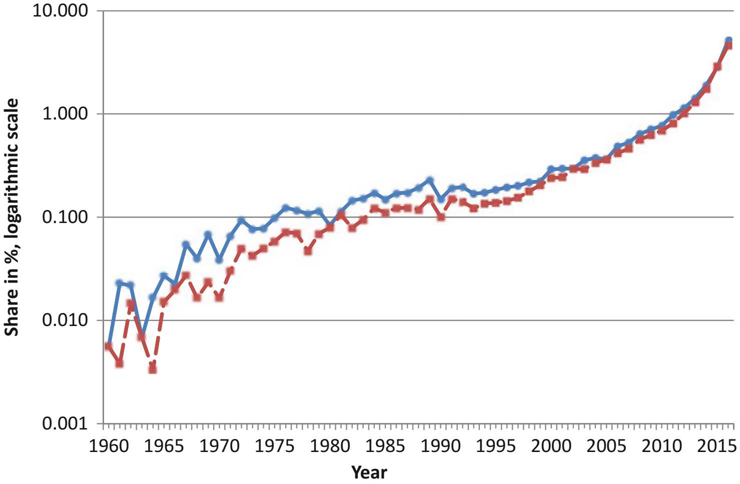


Fig. 1 Percentage citations of 'counterfactual' and of 'counterfactual impact evaluation' in economics. The percentages shown on the vertical axis are 100 *a*/*c* (solid line) and 100 *b*/*c* (dashed line), where *a*, *b* and *c* are the number of hits for the queries 'counterfactual AND economics', 'counterfactual impact evaluation AND economics' and 'economics', respectively

with 'economics' illustrates the increased incidence of counterfactual methods within economics (Fig. 1).

There is still a debate internal to economics on how reliable quasi-experimental methods are with respect to controlled experiments, which use randomisation (Bertrand et al. 2004). This is partly reflected in the idea of classifying studies according to 'strength of evidence' using the Maryland scale, which is often used to evaluate the evidence in justice cases (Sherman et al. 1997, 1998), and its modification by the UK What Works Network (Madaleno and Waights 2015).

In the EU policy context, and in particular in the assessment of regulatory policies, counterfactual impact evaluation methods are still rather novel. Limitations, heterogeneity of study designs and differences across areas and countries in administrative data access are factors that have limited the speed at which these methods have been introduced into the policy cycle.

To imagine how methods and policy applications might develop, it is useful to extrapolate trends in data gathering, access, linking, storage and availability. To this end, the following simplifying assumptions are made:

1. Microdata will cover all economic and social fields, and will be instantaneously updated.
2. Datasets will be linked, anonymised and made available to the research community.
3. (Unique) identifiers will allow seamless data linkage across policy areas (e.g. health, education, taxation).

4. Personal data protection will be fully ensured through effective legislation and oversight.
5. Micro-datasets will be available internationally in a fully comparable manner (e.g. pan-EU).

These assumptions may not necessarily appear realistic at present; however, they provide a useful framework for exploring policy opportunities and risks.

3 Stylised Aspects of a World with Perfect Administrative Microdata Availability

This section explores the implications of more, better and faster data in shaping policies, under the above assumptions (1. to 5.) reflecting the continuation of existing trends.

3.1 Breaking Silos: Multidimensional and Integrated Policy Design

Social and economic reality is shaped by the complex interplay of many factors reflecting numerous causal relations due to the actions of individual agents and groups. To a very significant degree, many policy-makers think in terms of isolated sectors, i.e. ‘in silos’. Here, an integrated approach, trying to understand the entire economic and social structure of an economy across sectors, can be contrasted with a pragmatic one, attempting to carry out intelligent policy analysis at a sectoral level.

The ‘silo culture’ is in no small part due to the data and information that policy-makers typically receive. Officials working on health policy, for example, generally read reports providing information and analysis on the workings of the health system for which they are responsible. Similarly, education departments tend to look at data from schools and universities, and tax collectors focus on company accounts and personal tax returns. The data used to evaluate these organisations, in turn, are based on distinct ‘health’, ‘education’ and ‘revenue’ metrics, respectively, which creates incentives to maintain the silo approach to policy.

Figure 2 shows this diagrammatically, using three traditional areas of public policy: education, health and tax collection. It exemplifies some arbitrarily chosen interlinkages between relevant variables in each silo. (Each link is represented by a double-headed arrow, even when there can be a one-sided causal link; this is to simplify exposition.) For example, it is not *prima facie* controversial to assume

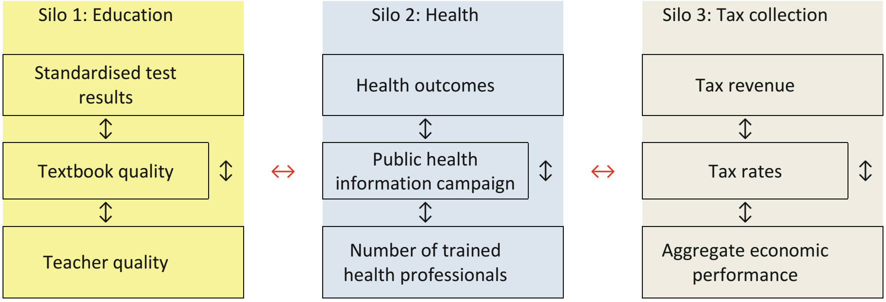

Silo 1: Education
Standardised test results
Textbook quality
Teacher quality
Silo 2: Health
Health outcomes
Public health information campaign
Number of trained health professionals
Silo 3: Tax collection
Tax revenue
Tax rates
Aggregate economic performance

Fig. 2 Vertical and horizontal policy links. A stylised example with three silos: education, health and tax collection. Links are represented by double-headed arrows (even when there can be a one-directional causal link) to simplify exposition. Vertical arrows are within-silo links and horizontal arrows represent links across silos

the existence of causal links between standardised test results, textbook quality and teacher quality. For instance, teacher quality and textbook quality may have an effect on standardised test results. Tax revenue depends significantly upon taxation rates and aggregate economic performance. These vertical linkages represent the within-silo links. Moreover, links can exist also across silos. For instance, overall tax revenue is important for funding expenditure on health and education; these links are indicated by the horizontal arrows.

The vertical, within-silo stratification of policy silo thinking is legitimate and has certain advantages. For example, while there are certainly interactions between education policy and health outcomes, the process of trying to quantify the linkages can be difficult. In the absence of abundant information about how different fields of government policy interrelate, there is a certain wisdom in focusing one's analysis in an area in which one has a certain level of expertise and understanding. However, under the assumption of perfect microdata availability, the silo mentality might be expected to attenuate over time. Subject to their availability, microdata that are easily linked across policy areas can potentially facilitate research into such cross-silo linkages.

One of the most important impacts of more readily available micro-datasets might then be a cultural shift in terms of how policy is designed and evaluated. If looking at data from the citizens' perspective—rather than the policy perspective—were to become predominant, then one might expect a more integrated approach to policy-making to develop over time.

While the speed at which changes in policy-making will occur is uncertain, the direction of the impact of greater availability of microdata is relatively clear. More integrated policy design should be the result, and it will hopefully drive better targeted and more effective public policy interventions. Indeed, such a 'cross-silo' approach is advocated by Gluckman (2017).

3.2 Ever More Precise Proxies for Citizen Well-Being

While the ultimate goal of public policy should surely be that of enhancing human well-being, data availability has been a major constraint on measuring the impact of government policies throughout history.[7] For example, in spite of a growing body of evidence that happiness and well-being are far from perfectly associated with wealth beyond some modest thresholds (Layard 2005), possibly due to the lack of better statistical proxies, government policies have focused mostly on measurable monetary and aggregate growth objectives.

Complete and linked microdata should allow better setting of policy objectives and targets, better policy design and better measurement of policy outcomes in terms of the ultimate objective of citizen well-being. Better proxies for well-being can first be determined on a range of factors including health, social and economic engagement, access to green areas, etc.

3.3 Reducing Evaluation Lags in the Policy Cycle and Adjusting Policies 'On the Go'

At present, both spending and regulatory policy evaluations are often available only many years after implementation. In many cases, quantitative evidence is missing, or based on ex post surveys, which are subject to well-known shortcomings. Where rigorous causal evaluation is undertaken, counterfactual impact evaluation methods are used. In practice, in the felicitous cases where good data are available, the most time-intensive part of this work is usually associated with obtaining the relevant datasets, and cleaning and linking the data across data sources, leaving very little time for analysis.

Subsequent econometric analysis, report writing and review procedures are also time consuming. With perfectly available microdata, the time lag from implementation to evaluation ought to be massively reduced. An evaluation strategy could be designed upfront, such that results would be available shortly after implementation. Depending on the type of policy, this could sometimes even be in near-real time.

This would further allow policy-makers to design policies conditional on intended outcomes (both on desired targets and on side effects), such that policy adjustment and fine-tuning could—within limits—become semi-automatic after a given probation or review period. *De facto*, in many cases, monitoring

[7] 'Whatever the form or Constitution of Government may be, it ought to have no other object than the general happiness' (Thomas Paine, *The Rights of Man*). The US Declaration of Independence of 1776 specifies 'the pursuit of happiness' as one of the principal inalienable rights of all citizens. The 1942 Beveridge Report in the United Kingdom, which laid the groundwork for Britain's welfare state, referred to 'the happiness of the common man' as the basic objective. The above citations are taken from Duncan (2010).

and evaluation frameworks could be merged. All this would have profound consequences for the duration of the policy cycle, and the effectiveness and efficiency of policies.

3.4 Reducing Other Lags in the Policy Cycle

3.4.1 Need for Policy

At present, the process of problem (or needs) identification in public policy typically arises from a government or institutional agenda, and/or from popular concern as expressed in the media and/or by civil society groups. Where an administration is tasked with designing in detail a policy measure to respond to this demand for action, the first phase is typically problem identification. This might take in the order of 1–2 years, to pull together the necessary data and analyse them and their evolution over time. With perfectly available microdata, this 'outside lag' ought to be cut dramatically, and perhaps to as little as a few months.

There would, in theory at least, be far less scope for controversy—and therefore far less need for analytical refinement—if microdata could instantaneously deliver a 'clear and transparent picture' of the *status quo ante* in a particular field of policy.

Moreover, open (but privacy-safeguarding) access to linked data for policy researchers could even trigger crowdsourcing of such analysis. For example, experimental opening of a linked labour market dataset at the IAB has led to a large number of policy research papers of the highest quality. This has placed the IAB in the top 6% of economic institutions as of September 2017 (for more details, see Chap. 7).[8]

However, if the use of such techniques is to become more widespread, they will have to move in step with measures that respond to citizens' fully legitimate concerns about personal data privacy and security, which are discussed in Sect. 4.

3.4.2 Policy Design and Consultation

At present, policy design and consultation for policy areas within the competence of the EU are the subject of a structured process implementing the Better Regulation Agenda.[9] Feedback is sought from interested stakeholders and impact assessments are prepared. As a rule of thumb, this process might typically take 1–2 years. Policy design and consultation processes are fundamentally human, democratic interactions and should certainly not be made subject to full automation simply because of the instantaneous availability of microdata.

[8] See https://ideas.repec.org/top/top.inst.all.html.

[9] See https://ec.europa.eu/info/law/law-making-process/better-regulation-why-and-how_en.

However, if such data were to become more widely available, there would be great potential to foster a better informed level of policy debate; see Spiegelhalter (2017). Robust evidence on policy impact would become more available, and at a more rapid rate. Enhanced knowledge about what works and for whom gained from fully linked data should facilitate enhanced policy research. The consultation and engagement process could then focus more on impacts that have been overlooked in past analyses. This 'inside lag' in policy design and consultation could therefore be shortened, to some degree, by better informed debates.

For the sake of completeness, as regards the legislative process, the main benefit would be through better informed discussions based on more robust evidence, rather than any significant time saving. Democratic due process would not necessarily be sped up by the availability of microdata.

3.5 The Potential of Administrative Data for Better Policy

Under the given assumptions, one would therefore expect to see a very considerable reduction in time lags in the policy cycle, from the identification of a policy problem to evaluating the policy's impact. Moreover, the availability of real-time administrative microdata would probably encourage processes of near-real-time policy adaptation. For example, one might imagine that policy redesign might be directly incorporated into some sort of dynamic evaluation process even during the course of the implementation phase. In democratic systems of governance, the challenge here is to ensure that more rapid policy analysis and adaptation, through improved administrative data availability, foster better informed policy design and consultation procedures. In this way, policy legitimacy can be ensured.

Summarising, both spending and regulatory policies could benefit from administrative data in a number of ways: (1) breaking silos to better incorporate interactions across policy areas; (2) allowing policy decisions and adaptation based on more robust evidence on what works for whom; (3) better measuring of impacts at individual or disaggregated group levels, reflecting distributions across income, age, ethnicity, location, etc.; and (4) much more efficiently and effectively targeting policy's ultimate objective of increasing citizen well-being.

4 Avoiding the Pitfalls

The potential benefits of perfect microdata for policy-making are great. Progress towards perfect microdata for policy could take policy-making to a different level. However, several threats are evident from both public and private sources. A number of obvious risks are considered in turn below, indicating how they could best be mitigated.

4.1 Data Privacy and Security

Perhaps the most obvious threat comes precisely from the risk of its de-anonymisation of microdata. Perfectly linked microdata allow the researcher to create an entire life narrative according to an individual's full set of recorded interactions with the organs of the state, which is, from a research point of view, a goldmine. From a privacy point of view, it is a potential danger, both at a personal and at governmental level; the concern is that data could be de-anonymised and released into the public domain. On this, all democratic countries have developed some level of right to privacy in their legal structure. This is certainly the case in the EU, where personal data protection has been significantly enhanced in recent years.[10]

Data security is an important issue around the globe. The fact is that, in contrast to the stylised assumptions in this chapter, data protection standards vary across different parts of the globe, as do levels of awareness among individual citizens about how to protect themselves from data theft. Moreover, hacks and data 'leaks' are a daily reality,[11] and some governments are themselves in the business of seeking to obtain personal data from other countries by illicit means.

Against this backdrop, citizens of any jurisdiction may not be entirely comfortable with perfectly linked and perfectly anonymised datasets being widely available to researchers around the world. Safeguards can, however, be implemented in the short term, along the following lines: (a) access to linked datasets can be limited to authorised researchers and officials; (b) work with those datasets can be restricted to controlled environments such as secure remote access, rather than the open Internet; and (c) access to data can further be limited to researchers and institutions that follow relevant procedures for democratic oversight and potential scrutiny of their activities.

These measures comply with the 'five safes' now common in this area: safe people (researchers can be trusted to use data appropriately), safe projects (research focuses on different populations of interest, and must be in the public interest), safe settings (data can be accessed in a secure environment), safe data (data is de-identified), safe output (research output must be confidentialised before it can be released); see Gendall et al. (2018).

Restricted access to microdata may in turn have implications for the development of open science and for academic peer-review procedures, as currently only a small number of researchers can replicate the work of selected peers. Moreover, international cooperation may also be curtailed to some extent by these data security concerns.[12] This appears to be one reason, for example, why some member states

[10] See http://ec.europa.eu/justice/data-protection/ and https://www.eugdpr.org/.

[11] See http://breachlevelindex.com/.

[12] The General Data Protection Regulation (GDPR) came into force in May 2018 and now sets the framework for the treatment of personal data throughout the EU; see https://www.eugdpr.org/.

of the EU are only at the earliest stages of discussions about granting access to non-national researchers.

4.2 Dislocation from Consultative and Legislative Due Process

As briefly discussed in Sect. 3, compressing many elements of the 'inside lag' in the currently conventional policy cycle will bring with it the challenge of ensuring that improved administrative data availability fosters better informed policy design and consultation procedures. As a result, some guidelines may become necessary to ensure enough time for society at large to engage with the policy implications brought about by faster policy research due to microdata availability.

4.3 Data Accuracy

An additional risk is simply a restatement of the well-known GIGO (garbage in, garbage out) phenomenon. Clearly, if linked micro-datasets became near-perfectly available and used in close-to-real-time policy adaptions, then any data imprecisions would be transmitted through the policy-making cycle at much higher speed. The implication is that even more attention will need to be paid to ensuring that data is as accurate as possible at the time of its being inputted into recording systems.

5 Concluding Remarks

The potential of using linked administrative microdata for better targeted policies in support of well-being appears to be very great, and possible associated pitfalls can be avoided. The likelihood of the assumptions on the availability to policy researchers of microdata being borne out in the future in Europe will depend on actions taken by public administrations, member states and the EU as a whole.

There are therefore strong arguments in favour of the increased use of administrative data to improve the quality and precision of impact evaluation and related public policy research. Use of public funds, such as EU strategic investment funds, could be envisaged for investment in this context. For this to happen, a shared set of objectives needs to be developed across the research, policy-making and data-holding communities. This will take time and will certainly need to take full account of concerns about data privacy and security.

Further refinements of the vision depicted in this chapter are desirable. Implementation of some of these ideas into legal frameworks and institutional processes will certainly require additional contributions from many stakeholders; this chapter attempts to get this discussion started.

Acknowledgements The authors thank Nuno Crato for useful comments on a previous version of this chapter and Francesco Tucci for assistance in the preparation of the graphs. The views expressed in this chapter do not necessarily reflect those of the European Commission.

References

Angrist JD, Pischke J-S (2017) Undergraduate econometrics instruction: through our classes, darkly. J Econ Perspect 31:125–144

Angrist J, Azoulay P, Ellison G et al (2017) Economic research evolves: fields and styles. Am Econ Rev 107:293–297

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? Q J Econ 119:249–275

Cisco (2017) The zettabyte era: trends and analysis. Cisco public white paper. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.pdf. Accessed June 2017

Duncan G (2010) Should happiness-maximization be the goal of government? J Happiness Stud 11:163–178

Feeney L, Bauman J, Chabrier J et al (2015) Using administrative data for randomized evaluations. J-PAL North America, Boston. https://www.povertyactionlab.org/sites/default/files/resources/2017.02.07-Admin-Data-Guide.pdf

Gendall K, McDowell A, Blackburn A (2018) Linking data for better policy: New Zealand's integrated data infrastructure. Stats NZ, Wellington, New Zealand

Gluckman P (2017) Using evidence to inform social policy: the role of citizen-based analytics. Office of the NZ prime minister's chief science advisor. www.pmcsa.org.nz/wp-content/uploads/17-06-19-Citizen-based-analytics.pdf

Hamermesh S (2013) Six decades of top economics publishing: who and how? J Econ Lit 51:162–172

Head BW (2015) Policy analysis: evidence based policy-making. In: International encyclopedia of the social and behavioral sciences, 2nd edn. Elsevier, New York, pp 281–287

Layard R (2005) Happiness: lessons from a new science. Penguin Books, London

Madaleno M, Waights S (2015) Guide to scoring methods using the Maryland scientific methods scale. What Works Centre for Local Economic Growth. http://www.whatworksgrowth.org/public/files/Scoring-Guide.pdf

Mastenbroek E, van Voorst S, Meuwese A (2015) Closing the regulatory cycle? A meta evaluation of ex-post legislative evaluations by the European Commission. J Eur Publ Policy 23:1329–1348

Panhans MT, Singleton JD (2017) The empirical economist's toolkit: from models to methods. Hist Polit Econ 49(Suppl):127–157. https://doi.org/10.1215/00182702-4166299

Sherman LW, Gottfredson DC, MacKenzie DL et al (1997) Preventing crime: what works, what doesn't, what's promising. National Institute of Justice. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.6206&rep=rep1&type=pdf

Sherman LW, Gottfredson DC, MacKenzie DL et al (1998) Summary. National Institute of Justice Research in Brief. https://www.ncjrs.gov/pdffiles/171676.PDF

Spiegelhalter D (2017) Trust in numbers. J R Stat Soc A 180:948–965

Stern E (2009) Evaluation policy in the EU and its institutions. N Dir Eval 123:67–85

Sven Langedijk is the Head of the Modelling, Indicators and Impact Evaluation Unit of the European Commission at the Joint Research Centre in Ispra (IT). The unit's research and analytical activities cover the full range of EU's policy areas, including education, competitiveness, employment, innovation, the internal market, judiciary, health, poverty, migration, cohesion, and other socio-economic dimensions. Previously, he headed the European Commission team responsible for developing, negotiating and monitoring the EU-IMF financial support and economic adjustment programme for Ireland. Until 2010, he contributed to the development of the EU fiscal governance framework and the Stability and Growth Pact and was editor of the Commission's annual flagship report on Public Finances in EMU.

Ian Vollbracht is the Deputy Head of the JRC unit Modelling, Indicators and Impact Evaluation. An economist and a researcher at the JRC, over the past 20 years he has worked in a variety of posts in the three main EU institutions: Commission, Council and Parliament. His work has mainly focused on economic, trade and budgetary policies. As a result of this varied career he has deep experience in the reality, as well as the theory, of policy development, implementation and evaluation within the EU institutions. Ian also has a deep interest in behavioural and psychological aspects of economics and policy-making. He is the co-author of the non-academic book *The Karmic Curve*, which seeks to bring some of these insights to a wider audience.

Paolo Paruolo is the Coordinator of the Competence Centre on Microeconomic Evaluation (CC-ME) at the European Commission—Joint Research Center, Ispra, IT. He has a Master in Economics and Statistics from the University of Bologna (1987) and a PhD in Mathematical Statistics (Theoretical Econometrics) from the University of Copenhagen (1995). He has taught Econometrics at the University of Bologna and at the University of Insubria (Varese, IT). His research interests are in econometrics (theory and practice) and in counterfactual methods. Because of his publication record, he was ranked among the best 150 econometricians worldwide in Baltagi (2007) "Worldwide econometrics rankings: 1989–2005", Econometric Theory 23, pp. 952–1012.