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 Policymaking 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.