Chapter 6
From Lack of Data to Data Unlocking

Computational and Statistical Issues in an Era of Unforeseeable Big Data Evolution

Nuno Crato

Abstract Reliable cross-section and longitudinal data at national and regional level are crucial for monitoring the evolution of a society. However, data now available have many new features that allow for much more than to just monitor large aggregates' evolution. Administrative data now collected has a degree of granularity that allows for causal analysis of policy measures. As a result, administrative data can support research, political decisions, and an increased public awareness of public spending. Unstructured big data, such as digital traces, provide even more information that could be put to good use. These new data is fraught with risks and challenges, but many of them are solvable. New statistical computational methods may be needed, but we already have many tools that can overcome most of the challenges and difficulties. We need political will and cooperation among the various agents. In this vein, this chapter discusses challenges and progress in the use of new data sources for policy causal research in social sciences, with a focus on economics. Its underlying concerns are the challenges and benefits of causal analysis for the effectiveness of policies. A first section lists some characteristics of the new available data and considers basic ethical perspectives. A second section discusses a few computational statistical issues on the light of recent experiences. A third section discusses the unforeseeable evolution of big data and raises a note of hope. A final section briefly concludes.

The author gratefully acknowledges the guiding questions and many helpful suggestions of Paolo Paruolo, as well as the constructive criticisms of Michele Vespe and the editors. The usual disclaimer applies.

N. Crato (✉)
ISEG, Cemapre, University of Lisbon, Lisbon, Portugal
e-mail: ncrato@iseg.ulisboa.pt

E. Bertoni et al. (eds.), *Handbook of Computational Social Science for Policy*,
https://doi.org/10.1007/978-3-031-16624-2_6

6.1 Introduction: Data for Causal Policy Analysis

A few decades ago, researchers and policymakers would struggle to get access to information. A student in time series would frequently have difficulty in getting data with 100 data points. A statistician willing to experiment with novel methods would frequently need to type data by hand, after collecting tables from dozens of print publications. An economist willing to compare the evolution of macroeconomic variables in different countries would need to search for days and would usually get series built with different criteria and with different length.

In the mid-1990s, things changed dramatically. Internet started working as an open means for communication and information access, although too many data sets were proprietary, as too many still are, and too often researchers would need to beg statistical officers or other researchers for getting appropriate data sets.

In parallel to an increasing data availability, a culture of openness spread slowly across countries and fields of activity. Driven by some governmental and institutional examples, by researcher pressure, and by public political tension, data that would previously be safely hidden in institution's departments become progressively available to researchers and the public.

Scientific journals could start avoiding systematically one of the obstacles to scientific reproducibility. Many journals adopted the policy of requiring authors to make data available upon request or by posting the data files at journals' websites.

In official statistics things started also changing. During the first years of the twenty-first century, the idea of using confidential microdata for research gained momentum (Jackson, 2019). This recent interest in original highly granular data officially collected, in brief, in administrative data, prompted the promise of a revolution in econometrics and social statistics studies.[1]

Microdata is usually defined as 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'.[2] Modern administrative data provides access to microdata at an unprecedented level.

This revolution in studies using administrative data was backed by a scientific "credibility revolution" in social statistics. Economists Angrist and Pischke (2015) described this "revolution" in empirical economics as the current "rise of a design-based approach that emphasizes the identification of causal effects". In fact, methods such as regression discontinuity, differences in differences, and others, which have been maturing in areas of statistical analysis as different as psychometrics or biometrics, registered a renewed interest as they become recognized as tools for assessing and isolating social variables influences and for looking for causal factors in overly complex environments. As already expressed in Crato and Paruolo (2019),

[1] For additional insights, please refer to the chapter by Signorelli et al. (2023) in the present Handbook.

[2] Glossary in Crato and Paruolo (2019, pp. 10–12).

this means that "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 policymaking can and should be increasingly supported by evidence".

By the end of the twentieth century, collected data volumes increased in such a way that researchers started using the phrase "big data". This phrase usually encompasses data sets with sizes beyond the ability of commonly used hardware and software tools to collect, manage, and process them within a reasonable time (Snijders et al., 2012). The expression encompasses unstructured, semi-structured, and structured data; however, the usual focus is on unstructured data (Dedić & Stanier, 2017).

Administrative data can be considered big data in volume, although usually it is highly structured and so it departs form this common characteristics of the big data classification.

This distinction is important as unstructured big data is evolving at an incredible speed, and it is by essence varied and difficult to characterize. What may be applicable to a big data set may not be applicable to a different big data set, and things are evolving at such a pace that new applications for big data are appearing every day. Very recently, the Covid-19 pandemic demonstrated the usefulness of new sources of data, such as students' logins to sites or the search for specific medical information. It will help our discussion to characterize the types of data we are discussing.

6.1.1 The Variety of Data

In this volume, the chapter by Manzan (2023) provides a valuable discussion of various sources of data and how they have been instrumental for advancements of knowledge in several fields of economics. Our purpose here is more schematic. In Table 6.1, we summarize the characteristics of different data types.

For our purposes, it is also interesting to characterize data according to their level of structuring. An attempt appears on Table 6.2.

For social research, policy design, and democratic public scrutiny, it is important to have access to as much data as possible, both in volume and variety. This is particularly important for data produced and kept by the public sector.

6.1.2 Underlying Statistical Issue: The Culture of Open Access

The idea that information should be available to the public is a democratic and an old one. The following well-known excerpt from James Madison, the father

Table 6.1 Types of data according to their origin, partially based on Connelly et al. (2016)

Origin → Characteristics ↓	General survey	Experimental survey	Administrative data	Other big data types
Research questions	Data addresses multiple questions	Data addresses specific questions	Data collected for non-research purposes	Data collected for non-research purposes
Structure	Highly systematic	Highly systematic	Systematicity varies	Very unsystematic
Dimensions	Large and complex	Reduced size and scope	Large and complex, but messy and fragmented	Very large and very complex
Sampling	Known sample and/or population	Known sample and/or population	Known sample and population	Unknown relationship sample population
Linkage	Difficult linkage	Linkage possible	Unique identifiers simplify linkage	Difficult linkage

Table 6.2 Types of data according to their structure (definitions and examples), loosely inspired by National Academies of Sciences, Engineering, and Medicine (2017)

Structured data from census and surveys	Structured public and private data	Semi-structured data	Unstructured data
Data from a population or a designed probability survey	Data collected by public administrations or from private companies	Data that have flexible structures that made it hard to relate them and need hard scrubbing and transformation for comparability	Data such as images, videos, and texts without any structure requiring value to be extracted and organized for processing and analysis
Examples			
Official censuses, academic and market research surveys, and other well-designed data collections	Tax records, school enrolments, unemployment, salaries, and other public records; commercial transactions, medical records, stock prices, and other private records	GPS and utility company sensors, tide and atmospheric sensors records, mobile texting volumes, web logs, web searches, and others	Internet searches, webcam traffic, security videos, medical data from personal sensors, social network interactions, and other data from IoT records (IoT: Internet of Things refers to devices that can communicate among themselves using the internet as the common transmission protocol)

of the American Constitution, has been recurrently quoted as an indictment of the withholding of government information (Doyle, 2022).

> A popular Government, without popular information, or the means of acquiring it, is but a Prologue to a Farce or a Tragedy; or, perhaps both. Knowledge will forever govern ignorance: And a people who mean to be their own Governors, must arm themselves with the power which knowledge gives.

More than two centuries later, similar concerns were clearly expressed in a report by President Obama's executive office (The White House, 2014), which 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).

In the European Union, there have been analogous concerns and recommendations. Among other statements, the European Commission has also pledged that, where appropriate, "information will be made more easily accessible" (2016, p. 5).

In addition to the issue of public access to nonconfidential data, there is the issue of data access for research purposes. This latter issue is an old one, but it took a completely different development in the twenty-first century with the rise of two factors: firstly, the availability of very rich, longitudinal, historically ordered, and granular administrative data; secondly, the development of the so-called counterfactual methods for detecting casual relations among complex social data.

In the United States, researcher's call to access to administrative data reached the National Science Foundation (Card et al., 2010; US Congress, 2016 ; The White House, 2014), which established a Commission on Evidence-Based Policymaking, with a composition involving (a) academic researchers and (b) experts on the protection of personally identifiable information and on data minimization.

Similar developments happened in Europe regarding the use of admin data for policy research purposes, albeit with heterogeneity across states. A few countries, namely, the UK and The Netherlands, already make considerable use of admin data for policy research. 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, aiming at better policymaking.

This research access has been discussed in general terms but has been dominated by policy concerns.[3] We are still far from regularly having the disclosure of administrative data and independent systematic analysis of policies. Too often, policy design is based on ideology, group interests, and particular policy matters,

[3] In science in general, the disclosure of scientific data and ideas has also benefited from the digitalization and the internet. The existence of scientific electronic archives that are nonrefereed and with open access, such as arxiv.org, and a variety of preprint archives is an open culture answer to the scientific priority concerns, making available data, experimental data, and ideas, is a way to establish priority (Watt, 2022).

without regard to its efficiency in terms of the intended goals. The possibility of measuring the impact of policies and correcting their course is certainly a very valid one and deserves all efforts for opening the access to data.

Although it is not clear whether this push for evidence-based policy impact evaluation is changing the panorama of policy design, it certainly is increasingly visible.

All these recent developments raise many questions and pose many opportunities and issues. In what follows, I will discuss three particular issues, trying to contribute to specific relevant policy questions raised by JRC scientists and collected by the editors of the volume in Bertoni et al. (2022). A first issue is how to take advantage of the different types of data by adding or consolidating the information available from each type of data set, ideally by linking them. A second issue is the scientific replicability of studies that access propriety data or data that evolves and are no longer retrievable. A third issue is confidentiality. With access to huge volumes of microdata, sensible personal or organizational information may be spread in a nonethical and undesired way. How can we navigate in this changing sea of opportunities without threatening legitimate privacy rights? These three main issues are tightly linked, as we can see in the following discussion.

6.2 Computational Statistical Issues

6.2.1 Statistical Issues with Merging Big Data

In contrast to organized administrative data, nonstructured or loosely structured big data are difficult to link with common probability linkage methods, namely, with those that are used to fix occasional misaligned units (Shlomo, 2019). There are, however, a few promising experiences.

A relatively old problem that can benefit from big data corrections is the so-called problem of the "missing rich", i.e. the paradoxical fact that too often data underestimates the size and wealth of people and families in the upper tail of the income distributions (Lustig, 2020). This has been a well-known problem in household surveys and other type of data collection in various countries.

The "missing rich" problem affects many types of data, not only in income distribution.[4] The expression now stands for issues that affect upper tails of social statistics, namely, underreporting, under covering and non-responses. For proceeding with estimates corrections, social statisticians have used methods that rely on within survey methods, looking for inconsistencies. More recently, there have been renewed interest in methods that rely on external sources, such as media lists and tax records. Researchers have used both parametric and nonparametric methods for these corrections. Corrections can be made by simple reweighing or

[4] See, e. g. Lustig (2020) and the references therein.

by adding items. In the first case, we are facing a trend to the use of model-based statistics, which have been common in areas as diverse as national statistics and student's standardized tests. In the second case, we are using selected administrative data linkage, as it has been done for a certain time in France for the EU-SILC survey.

Adamiak and Szyda (2021) work provide another example of merging official statistics with unstructured big data. They studied the distribution of worldwide tourism destinations by complementing the World Tourism Organization (UNWTO) data with two big data sources: a gridded population database and geo-referenced data on Airbnb accommodation offers. Their results emphasize the predominance of domestic tourism in the global tourism movements, an often-hidden phenomenon, which is revealed by a finer granular analysis of locations and types of tourism preferences. Global statistics with movements across borders cannot reveal the true scale of domestic movements.

Other researchers have explored similar big data sources for tracking dynamic changes in almost real time. For monitoring passenger fluxes, hotel stays, and car rentals, various researchers have successfully used booking data, Google searches, mobile device data, remote account logins, card payments, and other similar data. See, e.g. Napierała et al. (2020) and Gallego and Font (2021) as well as the work by Romanillos Arroyo and Moya-Gómez (2023) in the present Handbook.

Alsunaidi et al. (2021) provide a good synthesis of studies for tracking COVID-19 infections by using big data analysis. The pandemic prompted the surge of big data studies which were useful for estimation or prediction of risk score, healthcare decision-making, and pharmaceutical research and use estimation. Data sources for these studies have been incredibly varied, ranging from body sensors and wearable technology to location data for estimating the spread risks of COVID-19.

Additional data sources have been developed and should be most important in a foreseeable future. Among those, activity tracking and health monitoring through smart watches is proving to become an important tool. By using collected disperse data, researchers can now develop real-time diagnosis tools that could be used in the future. In his chapter in this volume, Manzan (2023) provides some other examples of microdata uses.

6.2.2 The Statistical Issue of Replicability and Data Security

The pandemic brought startling scientific advances in medicine and related areas but also in social statistics and in statistics in general.

A surprising reality that hit everybody was the uncertainty regarding many factors and variables in the pandemic. In early October 2020, the comparison of various estimates for the rate of Covid-19 spread in the United Kingdom revealed a degree of uncertainty masked by each individual estimate. Figure 6.1 shows the nine estimates considered at the time by the UK Scientific Pandemic Influenza Group on Modelling. The point estimates ranged from 1.2 to 1.5, i.e. widely different rates of

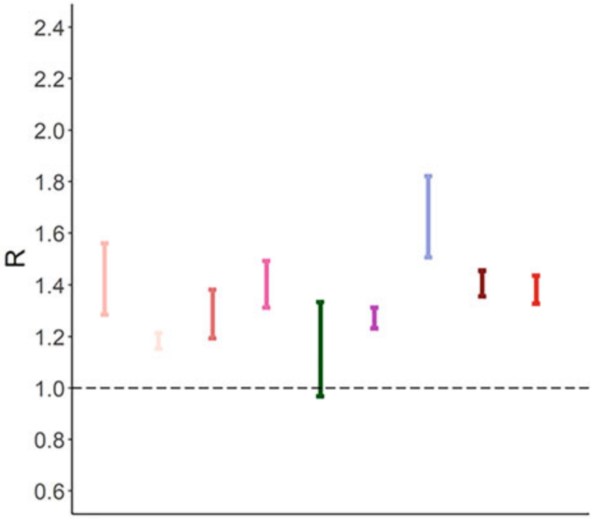

Fig. 6.1 Confidence bands at 90% for estimates for the reproduction rate R of Covid-19 in the UK in October 2022. Graph adapted from Scientific Pandemic Influenza Group on Modelling. (2020)

growth. Even more startling is the fact that different 90% confidence intervals do not overlap. The estimate represented as the fifth from the left on the graph admitted in the corresponding confidence interval the possibility that the pandemic is receding, while the highest estimate, the seventh on the graph, suggested that 100 people infect 166 others.

This example is not unique and similar results have recently been reported in other areas. A recent project in finance that collectively involved 164 teams tested six hypotheses widely discussed in financial economics (Menkveld et al., 2021). The hypotheses were on the existence of trends in the market efficiency, the realized bid-ask spread, the gross trading revenue of clients, and other measurable and testable characteristics of the markets. Additionally, used data were the same *Deutsche Boerse* sample.

Reporting the results from different teams, the authors note a sizeable dispersion in results. For the first hypothesis, for instance, which was that "market efficiency has not changed over time", the global standard error for the estimate was 20.6%, while the variability across researchers' estimates was 13.6%. This is certainly non-negligible.

The authors of this study propose to make a distinction between the traditional standard errors from parameters estimates, computed by using well-established statistical methods, and what they call "the non-standard errors", due to variability in methods used by researchers.

Along the same lines, a recent article in Nature (Wagenmakers et al., 2022) provides startling examples of different conclusions drawn from the same data with different statistical tools. Consequently, they argue persuasively on the need to contrast different research conclusions obtained through different statistical methods.

This would obviously be a particular form of triangulation, a concept worth revisiting.

Following the Oxford Bibliography by Drisko (2017), "triangulation in social science refers to efforts to corroborate or support the understanding of an experience, a meaning, or a process by using multiple sources or types of data, multiple methods of data collection, and/or multiple analytic or interpretive approaches". The concept was arguably first introduced by Campbell and Fiske (1959) and usually comprises four types of triangulation identified by Denzin in the 1970s: (1) data triangulation; (2) investigator triangulation; (3) theory triangulation; and (4) methodological or method triangulation.

As a way to apply triangulation and reaching more robust statistical conclusions in social sciences, Aczel et al. (2021) present a "consensus-based guidance" method and argue that a broad adoption of such "multi-analyst approach" can strengthen the robustness of results and conclusions in basic and applied research.

Wagenmakers et al. (2021) also argue that limitations of single analysis call for contrasting analyses and recommend seven concrete statistical procedures: (1) visualizing data; (2) quantifying inferential uncertainty; (3) assessing data preprocessing choices; (4) reporting multiple models; (5) involving multiple analysts; (6) interpreting results modestly; and (7) sharing data and code. For our purposes, this seventh recommendation is of paramount importance and consequences.

Let us highlight it again: for robustness of statistical inferences in social sciences, it is essential to share data and to share code. These have been practiced for decades in physical sciences. In particular, high-energy physics and astronomy have a long tradition of sharing data and procedures, so that other teams can replicate and corroborate, or contradict the analyses. A similar practice exists in climate research. Why is this such a novelty and odd thing to request in the social sciences?

A serious issue, though, is the security of sensitive data. Should data be completely free, easily available upon request, maybe entailing only a responsibility of a sworn statement, or should it be more rigorously restricted? There is no simple answer to this concern. But there are multiple practical solutions.

One practical solution is the availability to researchers of verified scripts only, with which studies could be done. This way, researchers do not deal directly with data and only get the statistical results. There are some inconveniences to this solution, namely, the difficulty in accessing data in this step-by-step way, while research usually needs to be done in an interactive way.

Another practical solution is the creation of safe environments in which only accredited researchers may have access and in which all interactions with data are recorded. With ethical and peer pressure from the scientific and technical community, this solution is feasible, although not without risks.

As a great provider of reliable data, public authorities should face in a very serious way the issue of safely organizing their data. A governmental example worth following is the X-Road, a centrally created and managed systematic data exchanger between information systems. It is extensively used in Estonia[5] and followed by Finland in 2017, when the exchange systems from both countries were interconnected.

6.2.3 Statistical Issues Risen by Anonymity Concerns and Related Challenges

Privacy is often quoted as the main concern for restricting the use of big data in various settings. This is obviously an important issue, but often shown through biased perspectives.

Firstly, it should be highlighted that tax collection, lack of respect for democratic rules in some countries, and the involuntary or unconscious supply of sensitive data to internet-based companies provide a much higher anonymity threat than big data studies operated by researchers following ethical protocols.

Secondly, the anonymity issue is often a convenient political pretext for not collecting data, not revealing data, nor assessing the impact analyses of public policies.

Thirdly, and most importantly, there are now methods of anonymizing data and realizing studies that do not reveal any personal sensitive data but provide the public with important knowledge about social issues.

Other issues are worth noting, namely, information correctness and replicability. Missing data and incorrect data can lead to biased findings (Richardson et al., 2020). And these incorrect findings can be replicated and induce larger mistakes. Additionally, data collected by businesses often change the sampling and processing methods and do not report it adequately (Vespe et al., 2021). All these issues are even more serious as they mean that replicability is often difficult and so the scientific debate can be hindered.

As we discuss big data availability and issues, it is obligatory to note that a wealth of administrative data of great use and of technically easy access exists and should be available to researchers and interested citizen groups. In this regard, if there are difficulties, they could easily be removed with sufficient governance will.

Rossiter (2020) has noted that access to education data is essential for institutions accountability. This could hardly be overstated as education arguably is one of the most important public policies issues and education budgets are among the most important in any country. What is a stake is highly important for a country's future and for the taxpayer, and what is at stake is the use of substantial public resources.

Read and Atinc (2017) listed the availability of education administrative data in 133 low- and middle-income countries and noted that 61 of these have no available data and 43 have only data at the national level. Of the 29 countries that have desegregate data, they were most in non-machine reading format, and only 16 of

[5] https://e-estonia.com/solutions/interoperability-services/x-road/

these provide data from student assessment. The consequent limitation can hardly be overstated: student results are the most—some can even say the only—important data regarding any education system.

This "underutilization of administrative data" has serious consequences form educational development. As Rossiter (2020) again points out, for many educational decisions findings cannot be imported. When there are conflicting evidence results, in particular, then "non-experimental results from the right context are very often a better guide to policy than experimental results from elsewhere".

We should thus look for solutions.

How can we replicate results if data are confidential and restricted to particular groups of researchers? We can address this issue by fostering communities of practice. This way, access to confidential data is guaranteed to trusted researchers under appropriate conditions. This would allow and nudge researchers to independently study the same data set and contrast conclusions.

Public and statistical authorities are among those more reticent to this type of data sharing. However, this is the best way to reach robust conclusions that can illuminate policy evaluation and public policy decisions.

In case a team of researchers claims that policy X had effect Y, one could ask a team of "research team of verifiers" to replicate or reanalyse the data to validate findings, similarly with what happens in physical sciences.

The "research team of verifiers" could even be reimbursed, as they provide a public service. But this could be done in exchange of similar work done by others (reciprocity), or as normal peer review work, which is often done for free.

In an ideal future, access to non-public administrative data could be regimented in a way that forces varied teams access and varied methods. This happens in public tenders. Why should not data access be granted mandatorily to more than a single research team? This prerequisite for data use would foster social sciences, public policy evaluation, and, ultimately, democracy. Publicly collected data is a public good.

A good example to this practice is what has been put in place by some scientific societies and scientific journals[6]: Data sharing is a requirement for paper publication.

A simple proposal is as follows. Similarly to what happens in scientific journals, official analysis of policies impact could have as a normal prerequisite the verification by independent researchers. In these cases, the analyses could involve much more computational and teamwork than normal paper refereeing. It would be of public interest that the promoter of the study includes in the initial budget a provision for paying teams of verifiers that could constitute an accredited pool.

[6] See, e. g. Committee on Professional Ethics of the American Statistical Association (2018).

6.3 The Way Forward

As discussed in Callegaro and Yang (2018) inter alia, variability is an important characteristic of big data. This means that gathering, analysing, and interpreting big data requires technical expertise that is always evolving. This also means that methods are evolving, and it is difficult or even impossible to have a fixed set of tools that will allow the use and merging of data, when we deal with this particular type of data.

Researchers have used relatively old or, at least, well-established techniques such as propensity score analysis, regression discontinuity, and differences-in-differences methods.

Another research worth noting is Chen et al. (2020). The authors note that the "challenge of low participation rates and the ever-increasing costs for conducting surveys using probability sampling methods, coupled with technology advances, has resulted in a shift of paradigm". At this moment, even government statistical agencies need to pay attention to non-probability survey samples, i.e. samples that are not random or that do not derive from a known probabilistic rule. One example is the so-called opt-in panels, for which volunteers are recruited. These authors propose a general framework for statistical inferences with this type of samples, by coupling them with auxiliary information available from a reference probability sample survey. In this setting, they propose a novel procedure for the estimation of propensity scores. All their procedure supposes the availability of high-quality probability sample surveys to allow for the pairing.

At this moment, data sources are evolving at such a speedy pace that it is difficult or even impossible to establish general rules. Each data collection method is providing new types of data with different characteristics, different insufficiencies, different challenges, and different possibilities. The general rules we may offer are (1) to apply established scientific rules and methods to the analysis of data and (2) to cross validate conclusions through open science, namely, through data and code sharing.

Is this a pessimistic or an optimistic view? I think it is an optimistic one.

6.4 Conclusion

This chapter discussed the recent evolution of data existence and use. It contrasted the previous lack of data with the current big data moment, in which we are facing a new issue, the issue of unlocking the power of existing data.

There are many types of data that fall under the classification of big data. This distinction is important, as methods to access, analyse, and use these types of data are different according to data structure. However, more than a practical issue, the wide use of data by the society is an ethical imperative. As such, this chapter argues that it is our duty as researchers to contribute to find ways of overcoming the many existing obstacles to full use of data.

There are many technical issues with data use, from anonymity issues to inference issues. This chapter lists some recent experiences and argues that some well-established scientific practices can be extended to data use and analysis, particularly when data are used for causal inference on policy measures. This can be done without increasing risks to data use and adding benefits to the scientific quality of the analyses. Scientific social studies and society will be the great beneficiaries.

References

Aczel, B., Szaszi, B., Nilsonne, G., van den Akker, O. R., Albers, C. J., van Assen, M. A., Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., Bringmann, L. F., Busch, N. A., Caruyer, E., Cataldo, A. M., Cowan, N., Delios, A., van Dongen, N. N., Donkin, C., van Doorn, J. B., et al. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife, 10*, e72185. https://doi.org/10.7554/eLife.72185

Adamiak, C., & Szyda, B. (2021). Combining conventional statistics and big data to map global tourism destinations before Covid-19. *Journal of Travel Research, 004728752110514*. https://doi.org/10.1177/00472875211051418

Alsunaidi, S. J., Almuhaideb, A. M., Ibrahim, N. M., Shaikh, F. S., Alqudaihi, K. S., Alhaidari, F. A., Khan, I. U., Aslam, N., & Alshahrani, M. S. (2021). Applications of big data analytics to control COVID-19 pandemic. *Sensors, 21*(7), 2282. https://doi.org/10.3390/s21072282

American Statistical Association. (2018). Ethical guidelines for statistical practice prepared by the Committee on Professional Ethics of the American Statistical Association approved by the ASA Board in April 2016. http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering metrics: The path from cause to effect*. Princeton University Press.

Bertoni, E., Fontana, M., Gabrielli, L., Signorelli, S., & Vespe, M. (Eds). (2022). *Mapping the demand side of computational social science for policy*. EUR 31017 EN, Luxembourg, Publication Office of the European Union. ISBN 978-92-76-49358-7, https://doi.org/10.2760/901622

Callegaro, M., & Yang, Y. (2018). The role of surveys in the era of "big data". In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research* (pp. 175–192). Springer International Publishing. https://doi.org/10.1007/978-3-319-54395-6_23

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Card, D. E., Chetty, R., Feldstein, M. S., & Saez, E. (2010). Expanding access to administrative data for research in the United States. *SSRN Electronic Journal.*https://doi.org/10.2139/ssrn.1888586

Chen, Y., Li, P., & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association, 115*(532), 2011–2021. https://doi.org/10.1080/01621459.2019.1677241

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

Crato, N., & Paruolo, P. (2019). The power of microdata: An introduction. In N. Crato & P. Paruolo (Eds.), *Data-driven policy impact evaluation* (pp. 1–14). Springer International Publishing. https://doi.org/10.1007/978-3-319-78461-8_1

Dedić, N., & Stanier, C. (2017). Towards differentiating business intelligence, big data, data analytics and knowledge discovery. In F. Piazolo, V. Geist, L. Brehm, & R. Schmidt (Eds.), *Innovations in enterprise information systems management and Engineering* (Vol. 285, pp. 114–122). Springer International Publishing. https://doi.org/10.1007/978-3-319-58801-8_10

Doyle, M. (2022). Misquoting Madison. *Legal Affairs*, July/August. https://www.legalaffairs.org/issues/July-August-2002/scene_doyle_julaug2002.msp

Drisko, J. (2017). *Triangulation [Data set]*. Oxford University Press. https://doi.org/10.1093/obo/9780195389678-0045

European Commission. (2016). *Communication to the Commission 'data, information and knowledge management at the European Commission*.https://ec.europa.eu/info/publications/communication-data-information-and-knowledge-management-european-commission_en

Gallego, I., & Font, X. (2021). Changes in air passenger demand as a result of the COVID-19 crisis: Using big data to inform tourism policy. *Journal of Sustainable Tourism, 29*(9), 1470–1489. https://doi.org/10.1080/09669582.2020.1773476

Jackson, P. (2019). From 'intruders' to 'partners': The evolution of the relationship between the research community and sources of official administrative data. In N. Crato, & P. Paruolo (Eds), Data-driven policy impact evaluation. Springer. https://doi.org/10.1007/978-3-319-78461-8_2

Lustig, N. (2020). *The "Missing Rich" in household surveys: Causes and correction approaches* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/j23pn.

Manzan, S. (2023). Big data and computational social science for economic analysis and policy. In *Handbook of computational social science for policy*. Springer International publishing.

Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johanneson, M., Kirchler, M., Razen, M., Weitzel, U., Abad, D., Abudy, M., Adrian, T., Ait-Sahalia, Y., Akmansoy, O., Alcock, J., Alexeev, V., Aloosh, A., Amato, L., Amaya, D., Angel, J. J., et al. (2021). Non-Standard Errors. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3961574

Napierała, T., Leśniewska-Napierała, K., & Burski, R. (2020). Impact of geographic distribution of COVID-19 cases on hotels' performances: Case of Polish cities. *Sustainability, 12*(11), 4697. https://doi.org/10.3390/su12114697

National Academies of Sciences, Engineering, and Medicine. (2017). *Innovations in Federal statistics: Combining data sources while protecting privacy* (p. 24652). National Academies Press. https://doi.org/10.17226/24652

Read, L., & Atinc, T. M. (2017). Information for accountability: Transparency and citizen engagement for improved service delivery in education systems. *Brookings Working Paper*, *99*. https://www.brookings.edu/wp-content/uploads/2017/01/global_20170125_information_for_accountability.pdf

Richardson, S., Hirsch, J. S., Narasimhan, M., Crawford, J. M., McGinn, T., Davidson, K. W., the Northwell COVID-19 Research Consortium, Barnaby, D. P., Becker, L. B., Chelico, J. D., Cohen, S. L., Cookingham, J., Coppa, K., Diefenbach, M. A., Dominello, A. J., Duer-Hefele, J., Falzon, L., Gitlin, J., Hajizadeh, N., et al. (2020). Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with Covid-19 in the New York City area. *JAMA, 323*(20), 2052. https://doi.org/10.1001/jama.2020.6775

Romanillos Arroyo, G., & Moya-Gómez, B. (2023). New data and computational methods opportunities to enhance the knowledge base of tourism. In *Handbook of computational social science for policy*. Springer International Publishing.

Rossiter, J. (2020). *Link it, open it, use it CDG note*. https://www.cgdev.org/publication/link-it-open-it-use-it-changing-how-education-data-are-used-generate-ideas

Shlomo, N. (2019). Overview of data linkage methods for policy design and evaluation. In N. Crato & P. Paruolo (Eds.), *Data-driven policy impact evaluation* (pp. 47–65). Springer International Publishing. https://doi.org/10.1007/978-3-319-78461-8_4

Signorelli, S., Fontana, M., Gabrielli, L., & Vespe, M. (2023). Challenges for official statistics in the digital age. In *Handbook of computational social science for policy*. Springer.

Snijders, C., Matzat, U., & Reips, U.-D. (2012). 'Big data': Big gaps of knowledge in the field of internet science. *International Journal of Internet Science, 7*(1), 1–5.

The White House. (2014). Big data: Seizing opportunities, preserving values. *Executive Office of the President*.

US Congress. (2016). *Evidence-based policymaking commission act of 2016, H.R. 1831, 114th Congress*.

Vespe, M., Iacus, S. M., Santamaria, C., Sermi, F., & Spyratos, S. (2021). On the use of data from multiple mobile network operators in Europe to fight Covid-19. *Data & Policy, 3*, e8. https://doi.org/10.1017/dap.2021.9

Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., van Dongen, N., Hoekstra, R., Moreau, D., van Ravenzwaaij, D., Sluga, A., Stanke, F., Tendeiro, J., & Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour, 5*(11), 1473–1480. https://doi.org/10.1038/s41562-021-01211-8

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature, 605*(7910), 423–425. https://doi.org/10.1038/d41586-022-01332-8

Watt, F. (2022, April 22). If you want science to move forward, you have to share it. *EMBL.* https://www.embl.org/news/lab-matters/if-you-want-science-to-move-forward-you-have-to-share-it/#:~:text=In%20December%202021%2C%20EMBL%20announced, research%20across%20the%20life%20sciences