International Journal of Social Research Methodology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tsrm20

Local data and upstream reporting as sources of error in the administrative data undercount of Covid 19

Joshua K. Dubrow

To cite this article: Joshua K. Dubrow (2021): Local data and upstream reporting as sources of error in the administrative data undercount of Covid 19, International Journal of Social Research Methodology, DOI: 10.1080/13645579.2021.1909337

To link to this article: https://doi.org/10.1080/13645579.2021.1909337

INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY
https://doi.org/10.1080/13645579.2021.1909337

Local data and upstream reporting as sources of error in the administrative data undercount of Covid 19

Joshua K. Dubrow

Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw, Poland

ABSTRACT

The Covid 19 pandemic illuminates the role data has in public policy-making, i.e. datafication of society, and the importance of exploring the local sources of data to reveal errors in what has assuredly been from the beginning an undercount of cases and deaths. I note four interrelated error sources. The first two are common to any quantitative data collection project: (1) representation, measurement, and data processing; and (2) problems of data standardization from unequally resourced local and national data providers. Covid 19 casts a special light on (3) the possibility of government intervention in at least the public presentation of these data; and (4) human errors in the data chain caused by a stressful data collection environment. To identify errors, we should look to national pressures and the local contexts from which these data are collected and the upstream reporting process.

ARTICLE HISTORY
Received 29 December 2020
Accepted 24 March 2021

KEYWORDS
Data; error; Covid 19; aggregation; harmonization; datafication

Quantitative data collection and dissemination, as part of the datafication of society, are sources of social power (e.g. Milan, 2020; Fourcade & Johns, 2020; see also Kotliar, 2020). Whereas data can be a benefit, they also contain errors, and many of those errors come from within the data collection process (Loukissas, 2019; Sandefur & Glassman, 2015). This process is local: The data in datification have deep attachments to their place of origin (Loukissas, 2019). Whereas there are many and various sources of error (e.g. Biemer, 2010; Groen, 2012), to make good policy decisions, we should also reveal those local sources and how they manifest in the data that governments, academics, advocacy groups, media, and others use.

The Covid 19 pandemic accentuates the need to identify the local sources of quantitative data and how these data are reported upstream. The growth, stability, and decline in cases and mortality across nations and time are generally based on administrative data that various people and organizations collect, harmonize, and aggregate. Many data providers have made their data available in machine-readable form for public scrutiny with an intention to guide policy-makers and educate the public. These data are the empirical foundation for the decisions of individuals, institutions, and policy-makers (Misra et al., 2020). The pandemic thus illuminates the importance of revealing errors in what has been, from the beginning, assuredly an undercount of cases and deaths.

In this research note, I consider these issues to explore the major sources of error[1] in Covid 19 administrative data. A contribution of this note is its focus on an unfortunately and particularly overlooked source: Upstream reporting of cases and deaths from the many and varied local health organizations to levels further up the data chain. Countries with different data infrastructures have different capacities to collect information and turn them into data (Granda & Blasczyk, 2010; De

CONTACT Joshua K. Dubrow Jdubrow2000@gmail.com Institute of Philosophy and Sociology, Polish Academy of Sciences, 72 Nowy Swiat, Room 211, Warsaw 00-330, Poland

Jong, 2016). Whereas other data situations revealed that the developing world is particularly underresourced (Heeks & Shekhar, 2019; Misra et al., 2020), Covid 19 shows that Western countries have data infrastructure problems at the local level too.

Local knowledge and upstream reporting

Data producers have long asked data users to consider where the data come from. A pernicious source of error is at the most local level where the data are collected and reported upstream to the research designers, aggregators, and disseminators.

Fundamental to data collection is what many refer to as 'local knowledge.' Local knowledge means recognition of the contexts in which these data were collected. The 3MC framework advises that 'local knowledge can be critical to understanding cultural traditions and customs, possible limitations, and the feasibility of the research' (De Jong, 2016). Separately, Jensenius (2014) argues that 'local knowledge also gives insights into how large datasets are collected, where their weaknesses lie, and how to spot irregularities in the data' (402). Data have attachments to their place of origin, such that 'data practices rely on local knowledge as well as experience for meaningful interpretation and responsible use' (Loukissas, 2019, p. 53).

As locally sourced data are reported upstream, errors may emerge at any point in the data value chain (Gal & Rubinfeld, 2019, p. 746). There is cross-national and cross-level variation in laws and norms of data collection, organization, analysis, storage, and use (De Jong, 2016; Gal & Rubinfeld, 2019, pp. 746–747). Perversely, a politicized bureaucracy and incentives to misreport become additional sources of systematic error in administrative data (Boräng et al., 2018; Sandefur & Glassman, 2015). Higher-level authorities of the data infrastructure may initiate or contribute to these errors; they ignore the local context at their peril.

To understand the locality of data, we should map the data landscape in which these data are produced and disseminated.

A landscape of administrative data on Covid 19 cases and mortality

For Covid 19, the data on infected persons (cases) and mortality (deaths), or what can be called 'counts,' are widely available.[2] Data providers and aggregators include well-known organizations such as the European Centre for Disease Prevention and Control (ECDC), Global Health 50/50, Google 'Covid 19', Johns Hopkins University COVID Tracker, *Our World in Data, The New York Times*, Wikipedia, the World Health Organization (WHO), and Worldometer, to name a few. Covid 19 data collectors and aggregators rely on a wide range of outside sources such as public authorities (various national and subnational health authorities, including from press conferences and social media) and reports in mainstream and social media (Twitter, Facebook, and Telegram). In their textual description, however, they can be vague about where exactly their data come from, and, perhaps to assure the reader that the data are thorough, some boast about how many sources they have. The ECDC claims that ' . . . a team of epidemiologists screen up to 500 relevant sources to collect the latest figures for publication . . . '[3] Worldometer writes that they 'validate the data from an ever-growing list of over 5,000 sources.' Some organizations simply present aggregate Covid 19 data from their fellow data aggregators, and, thus, errors can be repeated in multiple platforms.

Importantly, Covid 19 counts are *under*-reported.[4] Although coronavirus conspiracists argue that there is a case and mortality overcount, there is no logic or evidence for that argument.[5]

Sources of error

Data collection of Covid 19 counts is difficult. At root, organizations depend on information provided by various and unequally resourced local and national data collectors that, in turn, received it from hospitals, labs, and other health organizations and medical authorities that depend

on professionals within those organizations to report on Covid 19 cases and mortality. The upstream reporting process varies by nation, and descriptions of upstream reporting, in English, that share details of this process, are rare.[6] These difficulties are potential and interrelated sources of error. I note four.

The first source is common to all quantitative data collection processes, and for Covid 19, they are errors related to representation (e.g. the data contain people who were tested), measurement (e.g. inadequate testing instruments), and data processing (see Slomczynski & Tomescu-Dubrow, 2018; Oleksiyenko et al., 2018). On processing errors, one reason for changes in the numbers of Covid 19 counts is the continual re-definition of 'what is a case' and what counts as a fatality due to Covid 19 (Azzopardi-Muscat et al., 2021, p. 638). Many of these errors may be difficult to identify, let alone to correct ex-post, which in turn can influence the accuracy of statistics derived from Covid 19 counts across time. For example, the COVID Tracking Project for the US reports that the data situation has gradually improved, but they still have 'State data quality grades'.[7]

Second, imagine the impact of local reporting agencies – the tens of thousands of hospitals with unequal economic development – on standardization. Standardization requires that the data from a variety of sources and at different levels of aggregation are similar enough for comparison across nations, within nations, and over time (Gal & Rubinfeld, 2019). Different reporting standards and methods beget judgment calls from both the collector and the aggregator.[8] The WHO notes that the health data problems that range from submission to use have existed for decades and, to ameliorate, they have recently issued a call for each country to establish 'a national data coordination mechanism' (Azzopardi-Muscat et al., 2021, p. 638).

The third source is also a possibility with all quantitative data collection, but because of the magnitude of Covid 19, it has been publicized: As data are attached to local sources (Jensenius, 2014; Loukissas, 2019), there will be discrepancies in the quality of data reporting, perhaps from national government interference (Boräng et al., 2018; Sandefur & Glassman, 2015). For example, in May 2020, *The New Yorker* reported about Iran:

> Soon, Iran became a global center of the coronavirus, with nearly seventy thousand reported cases and four thousand deaths. But the government maintained tight control over information; according to a leaked official document, the Revolutionary Guard ordered hospitals to hand over death tallies before releasing them to the public.[9]

This is a stark example. Government's possible interference in data reporting can be more subtle, such as in Russia and Brazil.[10] In these situations, it is not clear whether there are data collection problems or the data are presented in ways that arouse suspicion.

Fourth are the conditions of the work environment in which these data are produced, and this includes the structural problems imposed on the people who labor to produce the data. Covid 19 data collection and reporting worldwide began in a novel pandemic, and this massive process was not properly standardized within or between nations. As a result, the people collecting the data, and the systems in which they work, have been unusually stressed.

To illustrate, I reconstruct events in the US based on the reporting about exactly this issue. Whereas approaches to collecting Covid 19 data can differ across nations, and the US case has idiosyncrasies, this reconstruction invites others to investigate, and publish in English, how this process works elsewhere.

Upstream reporting of Covid 19 cases and deaths depends on armies of white collar workers whose job is to fill out the reports of Covid 19 from hospitals, labs, and other health agencies. They depend on lab workers who take the samples and deliver the results. These essential lab and data workers enjoy job security (for now) but are vulnerable to mental and physical anguish. As reported in the US, 'testing teams are grappling with burnout, repetitive-stress injuries and an overwhelming sense of doom.'[11] Under strain since March 2020, by December, a sizable number simply quit, and replacements had not kept pace. As some in the US complain about not-enough-tests, or too-slow-tests, the lab workers are worn down by both the work and the crush of societal criticism.

Upstream from the local labs and other health centers is the Centers for Disease Control and Prevention (CDC). In Spring 2020, the CDC admitted that, in their harmonization and aggregation of data, they combined serology tests for antibodies with diagnostic tests of active viral infection, a data situation that may have led to a slight overcount of the number of Americans tested for Covid 19. Blame for this mishap was attributed to too much pressure on too much work in too short span of time, an understandable situation that unfortunately led to poor decision-making: 'Epidemiologists, state health officials and a spokeswoman for the C.D.C. . . . attributed the flawed reporting system to confusion and fatigue in overworked state and local health departments that typically track infections – not tests – during outbreaks.'[12]

The speed required in the pandemic to standardize across all parts of the reporting system whose infrastructure cannot handle the load can cause problems in the production of timely and accurate data. The CDC had been a case in point, as its data infrastructure had

> antiquated data systems, many of which rely on information assembled by or shared with local health officials through phone calls, faxes and thousands of spreadsheets attached to emails. The data is not integrated, comprehensive or robust enough, with some exceptions, to depend on in real time.[13]

Covid 19 demonstrates how local contexts – the people and the systems – react to both local and national pressures. The sudden and voluminous data demands of Covid 19 shocked each nation's multilevel data infrastructures (Misra et al., 2020, p. 5). As nations discovered that the health data produced by local sources are vital to national security, some pursue a top-down policy to standardize data reportage (Azzopardi-Muscat et al., 2021). Local sources, severely stressed and unequally resourced, felt pressure to meet the data needs at their own level and the speedy standardization demands from the national level. Errors may ensue.

Conclusion

Local, national, and cross-national data on Covid 19 cases and deaths have led to a host of policy decisions – social distancing, economic redistribution, and whether and how to conduct elections, among other things – that have impacted the lives of billions of people. Covid 19 count data are thus a source of social power. And these data have errors.

I follow the call from Jensenius (2014), Loukissas (2019), and others to illuminate the local contexts in which quantitative data are collected and reported upstream. I do not assess the validity and reliability of Covid 19 counts, but I have identified some major sources of error that may lead to discrepancies across nations and time.

This research note highlights an entirely human-made source of error that is underemphasized in the literature on quantitative data collection: Data errors may be due to unequal infrastructures of hospitals, labs, and other organizations staffed with time- and social-pressured people who, due to systemic problems and simply fatigue, make mistakes that can introduce a series of minor errors in the data that they report upstream. These errors can accumulate and manifest in the data chain in ways that we do not understand if we do not look for them.

Whether and how data collectors and disseminators take local knowledge into account, as suggested, is often unreported. There have been improvements in the collection and reportage of Covid 19 counts as organizations at different levels of administration gained more experience with the process and learned from previous mistakes. As countries update their knowledge and redefine cases and mortality, it is not clear whether they will retrospectively change their data to reflect the new knowledge.

In the pursuit of data, errors occur. The unheralded data collectors at the local level should be praised (and economically compensated) for their efforts. Identifying error sources in publicly released Covid 19 datasets will impact the beneficial use of these data for academic research and public policy.

Notes

1. Error can be defined as 'a measure of the estimated difference between the observed or calculated value of a quantity and its true value' (Google). Here, I am interested in the difference between observed values of Covid 19 cases and mortality as collected by people and organizations and the true counts across nations and time. Deviations from the true counts are errors. Reasons for the deviations can be called 'the sources of error.'
2. These counts are not raw numbers, especially at the individual or hospital level for obvious ethical and privacy reasons. See the WHO's 'Joint Statement on Data Protection and Privacy in the COVID-19 Response' https://web.archive.org/web/20210129063642/https://www.who.int/news/item/19-11-2020-joint-statement-on-data-protection-and-privacy-in-the-covid-19-response
3. https://web.archive.org/web/20210207154831/https://www.ecdc.europa.eu/en/covid-19/data-collection
4. *The Economist*. 'Tracking Covid-19 excess deaths across countries.'https://web.archive.org/web/20210316170856/https://www.economist.com/graphic-detail/coronavirus-excess-deaths-tracker*The New York Times*. 'Tracking the True Toll of the Coronavirus Outbreak'https://web.archive.org/web/20201229084144/https://www.nytimes.com/interactive/2020/04/21/world/coronavirus-missing-deaths.html
5. *The Independent*. 'US death toll could be double official figure.'https://web.archive.org/save/https://www.independent.co.uk/news/world/americas/us-coronavirus-real-death-toll-covid-29-cases-a9504911.html
6. According to *Our World in Data*, there is a '. . .long reporting chain that exists between a new case and its inclusion in national or international statistics. The steps in this chain are different across countries. . .' They claim that the chain exists 'for many countries.'https://web.archive.org/web/20210211103954/https://ourworldindata.org/covid-cases?country=IND~USA~GBR~CAN~DEU~FRA
7. https://web.archive.org/web/20,210,113,090,746/https://covidtracking.com/about-data/state-grades/There was a similar project in India (Vasudevan et al., 2020).
8. *The New York Times*. 'We're Sharing Coronavirus Case Data for Every U.S. County.' https://web.archive.org/web/20210208155606/https://www.nytimes.com/article/coronavirus-county-data-us.html
9. *The New Yorker*. 'The Twilight of the Iranian Revolution' by Dexter Filkins 25 May 2020 https://web.archive.org/web/20210316172411/https://www.newyorker.com/magazine/2020/05/25/the-twilight-of-the-iranian-revolution
10. *Bloomberg News*. 'Experts Question Russian Data on Covid-19 Death Toll.' 13 May 2020. https://www.bloomberg.com/news/articles/2020-05-13/experts-question-russian-data-on-covid-19-death-toll*BBC News*. 'Moscow more than doubles city's Covid-19 death toll' 29 May 2020. https://web.archive.org/web/20200529002935/https://www.bbc.com/news/world-europe-52843976*The Guardian*. 'Brazil stops releasing Covid-19 death toll and wipes data from official site.' https://web.archive.org/web/20210316173712/https://www.theguardian.com/world/2020/jun/07/brazil-stops-releasing-covid-19-death-toll-and-wipes-data-from-official-site
11. *The New York Times*. 'Testing-Lab Workers Strain Under Demand.'https://web.archive.org/web/20210204163909/https://www.nytimes.com/2020/12/03/health/coronavirus-testing-labs-workers.html
12. *The New York Times*. 'C.D.C. Test Counting Error Leaves Epidemiologists "Really Baffled".' https://web.archive.org/web/20210207172655/https://www.nytimes.com/2020/05/22/us/politics/coronavirus-tests-cdc.html
13. *The New York Times* 'Built for This, C.D.C. Shows Flaws in Crisis.' https://web.archive.org/web/20200607180433/https://www.nytimes.com/2020/06/03/us/cdc-coronavirus.html

Acknowledgements

This material is based upon work supported by the National Science Foundation (PTE Federal award 1738502) and by the National Science Centre, Poland (2016/23/B/HS6/03916). I am thankful for the comments by Irina Tomescu-Dubrow, Francesco Sarracino, Malgorzata Mikucka, Kazimierz M. Slomczynski, and the journal editors, and for the knowledge created by the Survey Data Recycling project on aggregation and harmonization at IFiS PAN and The Ohio State University. A version of this research note appeared in Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences 2020, v. 6, n. 1.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Joshua K. Dubrow is a Professor of Sociology at the Institute of Philosophy and Sociology, Polish Academy of Sciences (IFiS PAN), PI of a National Science Centre, Poland grant (2016/23/B/HS6/03916), and a co-founder of Cross-national Studies: Interdisciplinary Research and Training program (consirt.osu.edu)

References

Azzopardi-Muscat, N., Hans Henri, P., Kluge, S. A., & Novillo-Ortiz, D. (2021). A call to strengthen data in response to COVID-19 and beyond. *Journal of the American Medical Informatics Association*, *28*(3), 638–639. https://doi.org/10.1111/gove.12283

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848. https://doi.org/10.1093/poq/nfq058

Boräng, F., Cornell, A., Grimes, M., & Schuster, C. (2018). Cooking the books: Bureaucratic politicization and policy knowledge. *Governance*, *31*(1), 7–26.

De Jong, J. (2016). *General considerations in the online 3MC cross-cultural survey guidelines*. https://ccsg.isr.umich.edu/chapters/data-collection/general-considerations/.

Fourcade, M., & Johns, F. (2020). Loops, ladders and links: The recursivity of social and machine learning. *Theory and Society*, *49*(5), 803–832. https://doi.org/10.1007/s11186-020-09409-x

Gal, M. S., & Rubinfeld, D. L. (2019). Data standardization. *New York University Law Review*, *94*, 737.

Granda, P., & Blasczyk, E.: Data harmonization. Cross-cultural survey guidelines. http://ccsg.isr.umich.edu/pdf/13DataHarmonizationNov2010.pdf (2010)

Groen, J. A. (2012). Sources of error in survey and administrative data: The importance of reporting procedures. *Journal of Official Statistics*, *28*(2), 2.

Heeks, R., & Shekhar, S. (2019). Datafication, development and marginalised urban communities: An applied data justice framework. *Information, Communication & Society*, *22*(7), 992–1011. https://doi.org/10.1080/1369118X.2019.1599039

Jensenius, F. R. (2014). The fieldwork of quantitative data collection. *PS: Political Science & Politics*, *47*(2), 402–404.

Kotliar, D. M. (2020). Data orientalism: On the algorithmic construction of the non-Western other. *Theory and Society*, *49*(5), 919–939. https://doi.org/10.1007/s11186-020-09404-2

Loukissas, Y. A. (2019). *All data are local: Thinking critically in a data-driven society*. The MIT Press.

Milan, S. (2020). Techno-solutionism and the standard human in the making of the COVID-19 pandemic. *Big Data & Society*, *7*(2), 205395172096678. https://doi.org/10.1177/2053951720966781

Misra, A., Schmidt, J., & Harrison, L. (2020, April 14). *Combating COVID-19 with data: What role for national statistical systems? Paris 21: Covid 19 Response Policy Brief*. https://web.archive.org/web/20210210154603/https://paris21.org/sites/default/files/inline-files/COVID_Policybrief_Full.pdf

Oleksiyenko, O., Wysmulek, I., & Vangeli, A. (2018). Identification of processing errors in cross-national surveys. In T. P. Johnson, B.-E. Pennell, I. A. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology: Multinational, multiregional and multicultural contexts (3MC)* (pp. 985–1010). John Wiley & Sons, Inc.

Sandefur, J., & Glassman, A. (2015). The political economy of bad data: Evidence from African survey and administrative statistics. *The Journal of Development Studies*, *51*(2), 116–132. https://doi.org/10.1080/00220388.2014.968138

Slomczynski, K. M., & Tomescu-Dubrow, I. (2018). Basic principles of survey data recycling. T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds..), *Advances in comparative survey methodology: Multinational, multiregional and multicultural contexts (3MC)*. Ch. 43 (pp. 937–962). Wiley Hoboken.

Vasudevan, V., Gnanasekaran, A., Sankar, V., Vasudevan, S. A., & Zou, J. (2020). Variation in COVID-19 data reporting across India: 6 months into the pandemic. *Journal of the Indian Institute of Science*, 1–8. (Springer). https://doi.org/10.1007/s41745-020-00188-z