

Improving Data Collection and Management

Leah R. Fowler, JD, University of Houston Law Center; Jessica L. Roberts, JD, University of Houston Law Center; Nicolas P. Terry, LLM, Indiana University Robert H. McKinney School of Law

SUMMARY. Data are fundamental to good public health policies and their implementation. However, the lifecycle of public health data (collection, analysis, and distribution) in response to COVID-19 was flawed. Public health data suffered from politicization, a lack of centralized leadership, and substandard governance. These flaws must be quickly corrected. That rebuilding process should also seek to improve disease surveillance by leveraging syndromic surveillance, genomic surveillance, and digital epidemiology. Priority must also be given to addressing inequity by improving the amount and quality of sociodemographic data. As well as improving the quality of the data we collect, we must do more to make the data available to the parties that require it, presented in a form that maximizes its utility. Finally, our existing or novel institutions must find the appropriate balance between access and privacy.

Introduction

Sound public health policy and practice are evidence-based, driven by data that determine appropriate responses. For example, real-time information about who has a disease and where they live can help target interventions and resources, and provide valuable information about how a disease spreads within a population. When the data are inaccurate or incomplete, however, disease control measures suffer.

Many of the errors and missteps involving data collection and management during the first year of the COVID-19 pandemic in the United States were the product of politicization and inadequate leadership. Other data problems occurred even before data could be collected because public health agencies could not satisfactorily implement traditional contact tracing and digital app-based surveillance. There have also been sharp differences in the availability of testing for low-income and communities of color compared to more affluent areas home to largely insured, white people, further skewing the data collected and obscuring an unequal disease burden (Kim et al., 2020).

The pandemic exposed fundamental structural and data management flaws and the country's lack of an effective public health data system. Specifically, the United States lacks a unified structure for data gathering, management, and dissemination. But the errors that hindered pandemic response, such as politicization, lack of centralized leadership, and substandard data governance, also highlight a path forward. Improvement requires a uniform implementation of better models of disease surveillance and a concerted effort to identify and address inequity through targeted data collection. But, even the best data have limited utility if not

rapidly available to decision-makers. The distribution of useful data, be it more granular or in the aggregate, will require tailored data governance depending in significant part on both the types of data in a dataset and on its intended end-users. Deep datasets containing sensitive and potentially personally identifiable information may require a data trust. However, for the quick dissemination of aggregate data, like for pandemic dashboards, too much infrastructure can be a hindrance. We begin by considering the impediments to effective data collection, management, and dissemination in the current pandemic. We then turn to how we can improve data collection and distribution. We end with our recommendations for the future.

Problems Identified during COVID-19

Three major, often overlapping data problems are politicization, a lack of centralized leadership, and defects in data management policies.

Politicization

During the first year of the pandemic, access to COVID-related data, like numbers of positive tests, of available hospital beds, and deaths, felt like a zero-sum game. Increasingly, motivated individuals weaponized data to cast actors, entities, and environments in favorable or unfavorable lights, and sway public opinion. Instrumentalizing data in this manner occurred at all levels of government, facilitated by a systemic lack of transparency.

Federally, there was considerable dislocation of the traditional data responsibilities of the Department of Health and Human Services (HHS), the Centers for Disease Control (CDC), and the White House. Specifically, reports surfaced of active interference by political

appointees in the publication of even “untouchable” data sources such as the CDC’s *Morbidity and Mortality Weekly Report*. One of the many examples was the White House’s insistence that officials delete language on the dangers of singing from the CDC’s guidance on the reopening of churches in May 2020. Further, even after the CDC had upgraded its hospital tracking system, HHS took over the process, installing a private contractor to perform the data collection and tracking, severely undermining hospital compliance and data accuracy (Bandler et al., 2020). Federal actors similarly compromised data dissemination. For example, in December 2020, the White House Coronavirus Task Force stopped sending its tailored data and recommendations to each state on a proactive basis (Klein, 2020).

Similar stories played out in some states, typically when their governors sought to minimize the risks of COVID-19 and justify more lenient public health mitigation strategies. For example, in Florida, Governor Ron DeSantis reportedly fired the Department of Health’s data dashboard manager after she initially refused to delete records showing positive cases at a time when the governor was arguing for reopening the state. Subsequently, the manager set up an independent dashboard providing granular data about Florida’s cases and deaths (the COVID Monitor).

Lack of Centralized Leadership

COVID-19 has exposed shortcomings in the federated model of public health data management. The CDC has not asserted a strong leadership role in data collection, standards, reporting, and dissemination, and the states have taken divergent paths (Davenport et al., 2020). As a result, the country lacks a national standard for the reporting of COVID-19 test data. For example, states differ as to whether they report PCR tests, antigen tests, or both. States also have made frequent changes in the manner and frequency with which they report data. There are major differences in the mechanics of how different data or data from different sources are reported. For instance, some laboratory test data are first reported to state and local authorities before being passed on to the CDC. Other data are sent directly to CDC, while hospital laboratories report directly to HHS. These data problems resurfaced during the initial months of the vaccine rollout amid reports of serious flaws in the interoperability of federal databases such as Operation Warp Speed’s Tiberius and CDC’s VTrckS.

Because of delays in implementing reliable state and CDC dashboards, increasingly reliance has been placed on dashboards curated by media organizations such as the *Washington Post* or research institutions such as the Institute for Health Metrics and Evaluation. Additional, non-governmental tools have appeared to track effective reproduction rates (Rt.live, 2021) and predict the risks associated with various events and activities (mycovidrisk).

Substandard Data Governance

Data governance encapsulates collection standards, quality, integrity, and security of data during its lifecycle. One report concluded, “Unlike many other countries such as Germany, Senegal, South Korea, and Uganda, the United States does not have standard, national data on the virus and its control. The

[United States] also lacks standards for state-, county-, and city-level public reporting of this life-and-death information” (Prevent Epidemics, 2020).

This approach to data governance is the product of dangerous levels of fragmentation across multiple dimensions. The most obvious is across administrative institutions, with responsibilities split among federal, state, and local agencies. Outside of the public arena, fragmentation occurs among private entities, often driven by proprietary interests that prevent data sharing between actors.

Relatively early in the pandemic, researchers recognized that data lacked granularity about key sociodemographic variables (Krieger et al., 2020), particularly race and ethnicity. There was also chronic underreporting (as low as 10%) of asymptomatic infections in the first months of the pandemic (Perkins et al., 2020). There is still no data-informed national plan to direct vaccines to neighborhoods bearing the largest burden of disease.

Beyond substance, COVID-19 exposed flaws in public health data processes. Too much data is captured in or transmitted in analog form (such as by fax). As cases surged during the winter months of 2020, health departments were often overwhelmed by the volume and logistics of processing testing data, the majority of which was not delivered digitally (Pearlstein & Moser, 2020). The resulting delay inhibited timely and targeted interventions.

COVID-19 data governance is overdue. Questions about indicators, such as whether “confirmed cases” include “presumptive positive cases” require standardized answers. Data are also fragmented by type or purpose. For example, demographic, racial and ethnic, clinical, and research data are viewed as distinct. Finally, like many aspects of health care, effective and efficient public health responses require collaboration and coordination between diverse groups, including providers, laboratories, and public health agencies. An individual may interact with the system at any of these points, and the ability to draw inferences requires connecting the dots. Improving data and data sources ultimately also requires a long-term investment in interoperability.

Improving COVID-19 Data

As noted above, fast and accurate data are critical for an effective and tailored public health response to a pandemic. However, data-driven interventions are only useful if the data underlying their design are reliable, high quality, and timely available. Several data categories should be part of mandatory pandemic reporting and made available to the public. This includes expanded surveillance approaches and data that help answer the who, what, when, where, and how of disease burden and spread. Tracking and addressing health disparities should be incorporated by design, with standardized reporting requirements for demographic information, congregate living, and secondary pandemic impacts like suicide and substance dependency.

Improving Disease Surveillance

While case counts are a key data point in pandemic response, they may lag behind broad community spread due to delays in test results and the onset of symptoms prompting an individual to seek

testing occurring after a patient is initially contagious. However, surveillance can identify community spread before it is indicated by clinical tests and hospitalization — a point by which early interventions are less effective. As a result, both biological and digital surveillance will be critical data sources for avoiding future waves of infection.

Biological Surveillance. Syndromic surveillance is a cornerstone of public health activity. It has long helped monitor flu, flu-like illnesses, and even potential bioterror attacks. Other indicators, particularly monitoring virus levels in sewage, are particularly useful for SARS-CoV-2. Research has shown sewage surveillance provides notice of community spread in advance of both hospitalizations and test result reports (Peccia et al., 2020). However, the lead time can vary depending on the speed with which localities can process and report test results (Peccia et al., 2020). These forms of surveillance take on increased importance in light of insufficient and inconsistent access to traditional tests and may provide enough early notice to slow community spread before cases overwhelm health care and public health systems.

Genomic surveillance has also been the key to understanding how COVID-19 has spread nationally and internationally. More specifically, understanding how and where outbreaks occurred in Germany and Washington State suggests that “intensive, community-level respiratory virus surveillance architectures” and genomic analysis are of particular value in reacting to future viruses (Worobey et al., 2020). Genomic surveillance is also essential for understanding mutations to the virus over time, helping identify potential changes in virulence and infectiousness. Reports suggest that the United States lags behind other countries such as the United Kingdom in collecting and analyzing virus samples.

Digital Epidemiology. Beyond the formal medical and public health infrastructure, digital epidemiology can improve detection and analysis. Digital epidemiology is a form of public health surveillance based on diverse data sources collected for non-public health purposes, such as mobile phone location data. Surveillance of internet searches and online activity can also predict an outbreak before more traditional mechanisms (Ginsberg et al., 2009). Other innovative forms of surveillance have proven particularly promising for the COVID-19 pandemic, both online and on the ground. Artificial intelligence, such as the BlueDot algorithm, famously identified early in the pandemic in December 2019, several days before the World Health Organization’s (WHO) announcement, by analyzing online activity.

Unlike more traditional public health surveillance, digital epidemiology presents unique challenges. Obstacles include privacy and access to proprietary data (Tarkoma et al., 2020). Scholars have argued that the benefits of disease forecasting or modeling, and sophisticated contact tracing may need to override individuals’ privacy interests. However, this should only occur when the alternatives — such as lockdowns — are worse. There should also be a responsible, transparent oversight process with broad representation from all stakeholders (Mello & Wang, 2020).

Addressing Inequity through Improved Data Collection

In addition to where the disease is spreading, it is critical to understand who bears the burden of disease and where and how they contract it. However, the collection of data on variables like race, ethnicity, income, and housing, or food insecurity has not been prioritized. By the end of 2020, only a handful of states reported COVID-19 testing data by race, limiting policymakers’ abilities to equitably allocate resources like testing, education, and support. The impact of COVID-19 on people with disabilities was also sorely lacking, in part due to the lack of accessibility of tests and testing centers and in part due to how data were collected (Reed et al., 2020). For example, drive-through testing sites exclude individuals who do not drive. Similar problems arose with vaccine distribution with many states failing to collect race and ethnicity data notwithstanding a CDC mandate.

It is also critical to understand disease distribution. Contagious diseases take the most significant toll on those who live in close proximity to others, such as in long-term care facilities, prisons, and detention centers. Even though these living conditions are most vulnerable to spread of COVID-19, states inconsistently collect and report data on cases, deaths, and locations, obscuring the burden’s true extent. Similar data collection deficiencies have hindered our understanding of the disease burden by occupation, including health care workers and employees in high-risk industries such as food processing.

Collecting and reporting these data are necessary for rapid pandemic response and contribute to the long-term understanding of the effects of that response on the population as a whole. For example, while nonpharmaceutical interventions like social distancing, isolation, and quarantine are essential tools to combat COVID-19, they also fracture social networks and support systems. Exacerbating this sudden loss of human connection is an environment of economic uncertainty and increased barriers to care (Reger et al., 2020). Social isolation is associated with worse health outcomes generally (Holt-Lunstad, 2017) and may lead to increases in cases of preventable death, like suicide (Reger et al., 2020). However, establishing these connections between secondary outcomes and pandemic interventions requires more and better data.

While data collection and reporting efforts by the media and other private actors are laudable, they are insufficient. Uniform policies and standards are sorely needed to capture these data to understand the burden of disease and to target limited resources to where they can have optimal impact. To do so requires a coordinated response, including a centralized, trusted agency in charge of data collection and evidence-based policy recommendations (Davenport et al., 2020). Some data can and should be collected, stored, and reported only in the aggregate. Some data must be more granular and identifiable to be useful. These datasets present different risks and challenges, and governance must be tailored to meet those needs.

Improving Data Distribution

An effective public health response requires that the right people can quickly access reliable information to make informed decisions. The United States botched its COVID-19 response in part because of serious missteps not only regarding data collection and management, but also its distribution. In addition to improving the quality of the data collected, we must ensure the data and derived information — once collected — are both secure and readily available to the parties that require them.

Both scientists and laypeople find dashboards, interactive online public health tools that provide community members with pandemic-related information in a given geographic area particularly useful. The CDC currently maintains a federal dashboard of data submitted to the agency (COVID-19 Module Data Dashboard). Other, extant dashboards provide data regarding states (e.g., Washington State Department of Health), counties (e.g., Harris County Public Health), nursing homes (AARP Public Policy Institute), and universities (e.g., Indiana University).

Pandemic dashboards should have a stated purpose — to provide reliable up-to-date, local, COVID-19-related information — and clear uniform policies about how they collect, manage, and protect their data. Best practices should be followed, and dashboard curators should work to standardize data presentations, for example whether to present data on a linear or logarithmic scale. The goal of pandemic dashboards is to provide citizens with reliable, up-to-date information about the pandemic in their area. Facilitating quick, easy access to accurate dashboard data is particularly important for older and other high-risk or vulnerable individuals so that they can make informed decisions.

The entities charged with warehousing data must strike the right balance between facilitating swift, straightforward data access to the proper stakeholders with ensuring privacy and security for sensitive information. One potentially useful model would be to establish a “data trust.” Data trusts gained popularity in the United Kingdom as a means for facilitating data sharing while protecting the rights of data sources. A data trust has five key elements: (1) compliance with all relevant legal standards in the given jurisdiction related to data collection, distribution, and management; (2) clear data governance structures; (3) well-defined data management processes and policies; (4) required trainings for data users; and (5) public and stakeholder engagement (Paprica et al., 2020).

In the wake of the pandemic, as the United States reconsiders the level of independence required for important agencies such as the CDC and Food and Drug Administration (FDA), consideration should also be given to establishing a public health data trust as an independent federal agency, potentially named the Federal Public Data Agency (PDA). The PDA would be charged with rulemaking related to data standards, governance, and protection.


Conclusion

Politicization, lack of centralized leadership, and substandard data governance hindered initial responses to the COVID-19 pandemic, but they need not remain stumbling blocks. Improving pandemic

response requires an intentional approach to data collection on both a macro and micro scale. Broader surveillance—in the traditional biomedical and public health sense as well as its novel, digital forms—can help policymakers stay ahead of the curve, obviating the need for controversial and disruptive control measures. Detailed, uniform data collection on key demographic variables can help decision-makers target limited resources intentionally to alleviate disparities in disease burden. But these approaches involve varying levels of risk and require different types of governance.

Ultimately, any sound data governance and distribution policy will depend in significant part on both the type of data in a dataset and on its intended end users. A rich dataset that includes comprehensive and potentially identifiable information requires more policies and safeguards than a pandemic dashboard that communicates only a single form of aggregated data. While the former is of use to researchers and public health authorities, the latter targets the general public. Data security and preventing unauthorized secondary use is important for potentially revealing datasets in the hands of sophisticated parties that might include the government and private companies. By contrast, ease of access is crucial when the dataset is limited, and the anticipated user is an ordinary citizen seeking to make an informed decision in real-time. Going forward, we must be careful to develop clear, transparent, flexible data governance structures tailored both to the kinds of data being collected and to the desired end users of that information.

The Biden administration clearly recognizes the country’s data challenges and one of the incoming president’s first executive orders ordered a sweeping review of the public health data infrastructure. At the federal level there must be one national agency charged with data collection. That agency must set the data standards for tests, cases, deaths, and sociodemographic data. The agency and its leadership must also “foster a data-driven culture” for future public health challenges (Davenport et al., 2020). A system cannot respond effectively to inequities in the absence of data. Data regarding race, ethnicity, income, and housing or food insecurity must be included in data sets and in analyses.

At the state level, all dashboards should adopt similar user interfaces and provide access to similar levels of granular data on a timely (daily) basis, including the 15 essential indicators. State dashboards also should follow best practices such as preferring rates over counts, smoothing data over time, “clearly identifying the intended audience, prioritizing key measures, having a clear organization and layout, presenting information to inform on health equity, updating information daily, and clearly labeling data and graphics” (Prevent Epidemics, 2020). 

Recommendations for Action

Federal government:

- The federal government should designate a single federal agency or data trust in charge of public health data with clear and transparent communications with state and local public health agencies to build trust.
- The federal government should charge that agency with establishing accountability and overseeing enforcement for inappropriate data use.
- Federal and state governments working together should improve disease surveillance by dramatically increasing syndromic surveillance, genomic surveillance, and digital epidemiology.
- The federal government should publish clear and transparent policies and processes based on scientific best practices for collecting, maintaining, and disseminating data.
- The federal government should standardize data types, collection and transmittal models through legislation, regulations, model statutes, or professional guidelines.
- The federal government should prioritize the collection of sociodemographic data particularly as it impacts disparities and health equity.
- The federal government and Congress should work with industry and other developers to ensure that the technologies used by the government adhere to the highest possible privacy and security standards.

State governments:

- States should adhere to existing laws, regulations, and best practices at both the federal and state levels for collecting, maintaining, and disseminating data.
- States should standardize state-, county-, and city-level public reporting using data standards consistent with federal standards.
- States should comply with CDC mandates on the collection of race and ethnicity data during vaccine distribution.
- States should create streamlined and transparent processes for disseminating up-to-date, actionable data (such as data dashboards) to citizens.
- States should engage citizens by making data readily accessible for public use (e.g., pandemic dashboards), educate the public regarding new research or developments, and solicit and respond to feedback regarding these resources.



About the Authors

Leah R. Fowler, JD, is a Research Assistant Professor and Research Director in the Health Law & Policy Institute at the University of Houston Law Center. She conducts multidisciplinary, collaborative research on the ethical, legal, and social implications of health care innovations, including consumer health technologies, novel care delivery models, and public health interventions. Her scholarship appears in social science, bioethics, and legal publications. Prior to UH, Professor Fowler was the Health Policy Program Manager at Baylor College of Medicine's Center for Medical Ethics & Health Policy, where she maintains a designation as a Health Policy Scholar.

Jessica L. Roberts, JD, is the Director of the Health Law & Policy Institute and the Leonard Childs Professor in Law at the University of Houston Law Center. She specializes in genetics and the law, health law, and disability law. Prior to UH, Professor Roberts was an Associate-in-Law at Columbia Law School and an Adjunct Professor of Disability Studies at the City University of New York. Immediately after law school, she clerked for the Honorable Dale Wainwright of the Texas Supreme Court and the Honorable Roger L. Gregory of the Fourth Circuit Court of Appeals. Professor Roberts has received the university-wide Teaching Excellence Award and the Provost's Certificate of Excellence. She was named a 2018 Greenwall Faculty Scholar in Bioethics and is a Health Policy Scholar with Baylor College of Medicine's Center for Medical Ethics & Health Policy.

Nicolas Terry, LL.M., is the Hall Render Professor of Law at Indiana University Robert H. McKinney School of Law where he serves as the Executive Director of the Hall Center for Law and Health and teaches various health care and health policy courses. His recent scholarship has dealt with health privacy, mobile health, Big Data, AI, and the opioid overdose epidemic. He has received several grants relating to substance use. In 2016, he testified before the House Energy and Commerce subcommittee on the Regulation of Mobile Health Apps and in 2018 before the Senate Committee on Aging on opioids policy. He blogs at Harvard Law School's *Bill of Health*, his "The Week in Health Law" podcast is at TWIHL.com, and he is @nicolasterry on twitter.

References

- Bandler, J., Callahan, P., Rotella, S., & Berg, K. (2020). Inside the Fall of the CDC. *ProPublica*.
- Davenport, T. H., Godfrey, A. B., & Redman, T. C. (2020). To Fight Pandemics, We Need Better Data. *MIT Sloan Management Review*.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012-14.
- Holt-Lunstad, J. (2017). The Potential Public Health Relevance of Social Isolation and Loneliness: Prevalence, Epidemiology, and Risk Factors. *Oxford Academic Public Policy & Aging Report*, 27(4), 127-30.
- Kim, S. R., Vann, M., Bronner, L., & Manthey, G. (2020). Which Cities Have The Biggest Racial Gaps In COVID-19 Testing Access? *FiveThirtyEight*.
- Klein, B. (2020). White House coronavirus task force no longer proactively sending reports to states. *CNN*.
- Krieger, N., Gonsalves, G., Bassett, M. T., Hanage, W., & Krumholz, H. M. (2020). The Fierce Urgency Of Now: Closing Glaring Gaps In US Surveillance Data On COVID-19. *Health Affairs Blog*.
- Mello, M. M. & Wang, C. J. (2020). Ethics and governance for digital disease surveillance. *Science*, 368(6494), 951-954.
- Paprica, P. A., Sutherland, E., Smith, A., Brudno, M., Cartagena, R. G., Crichlow, M., ... Yang, K. (2020). Essential Requirements for Establishing and Operating Data Trusts: Practical Guidance Based on A Working Meeting of Fifteen Canadian Organizations and Initiatives. *International Journal of Population Data Science*, 5(1), 31-40.
- Pearlstein, J. & Moser, W. (2020). Pandemic Data Are About to Go Sideways. *The Atlantic*.
- Peccia, J., Zulli, A., Brackney, D. E., Grubaugh, N. D., Kaplan, E. H., Casanovas-Massana, A., ... Omer, S. B. (2020). Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. *Nature Biotechnology* 38(10), 1164-67.
- Perkins, T. A., Cavany, S. M., Moore, S. M., Oidtman, R. J., Lerch, A., & Poterek, M. (2020). Estimating unobserved SARS-CoV-2 infections in the United States. *Proceedings of the National Academy of Sciences*, 117(36), 22597-602.
- Prevent Epidemics. (2020). Tracking COVID-19 in the United States From Information Catastrophe to Empowered Communities. Retrieved January 28, 2021, from <https://preventepidemics.org/wp-content/uploads/2020/07/Tracking-COVID-19-in-the-United-States-Report.pdf>
- Reed, N. S., Meeks, L. M., & Swenor, B. K. (2020). Disability and COVID-19: who counts depends on who is counted. *The Lancet Public Health*, 5(8), e423.
- Reger, M. A., Stanley, I. H., & Joiner, T. E. (2020). Suicide Mortality and Coronavirus Disease 2019—A Perfect Storm? *JAMA Psychiatry*, 77(11), 1093-94.
- Worobey, M., Pekar, J., Larsen, B. B., Nelson, M. I., Hill, V., Joy, J. B., ... Lemey, P. (2020). The emergence of SARS-CoV-2 in Europe and North America. *Science*, 370(6516), 564-70.