09

Health Data Infrastructure
The United States has been forced to rely on Britain and Israel to provide crucial insights about the Covid pandemic because a patchwork health data infrastructure leaves the country partly blind. To better protect American lives, jobs and communities, the federal government must spend money and effort to create a national health data platform with constant and consistent inputs from every state and territory.

A Tattered Quilt of Dislocated and Disorganized Data

During a pandemic, real-time data acquisition, analysis, and sharing is essential for informed decision-making. The current health data infrastructure is fragmented and underfunded, which has led to five primary challenges hindering effective public response. The first is a lack of a secure, standardized, and real-time national data platform for SARS-CoV-2 and other health threats. Second, there are significant variations in data quality and source from public and private entities. Third, insufficient linkages across data types hinder analysis. Fourth, the timing of reports varies and are often delayed. And fifth, data analyses are generally limited, non-transparent and delayed as well.

Epidemiological analysis, outcomes tracking, and vaccine safety reporting and effectiveness have increasingly come to rely on the United Kingdom and Israel (among other nations) in large part because their health data infrastructures are more comprehensive, reliable, and timelier than that of the United States.

The nation’s health data infrastructure needs a complete overhaul. In the meantime, accelerated efforts are needed to standardize, consolidate, and link data in the short term while preparing the ground for a thoroughgoing modernization. The bipartisan PREVENT Pandemics Act introduced in the United States Senate offers some hope that both can be accomplished. The Act calls for the CDC Director to develop a comprehensive plan for strengthening both the government’s internal systems and improving the bridges to private sector approaches, although it does not address the CDC’s lack of legal authority to collect Covid-related data after the public health emergency declaration expires, nor its lack of authority to require that states collect and report public health data to the federal government, both of which should be addressed in future congressional legislation.
How Things Got So Bad

A core feature of the nation’s health infrastructure is its decentralized structure, a legacy of federalism. Most of the funding comes from the federal government, and most of the data comes from the states. While empowering fifty state “laboratories” can be beneficial in identifying and evaluating good policy, the federalist model has been a disaster for efforts to identify and fight infectious diseases, contaminated food outbreaks and other national health threats because there is no centralized, real-time data platform for most public health data.

Much of the needed data are scattered across different and often incompatible computer systems.

To be sure, some states like Arizona have invested in a robust public health reporting infrastructure and do a good job tracking threats. But others like Florida have done little. This spasmodic approach prevents the collection of comprehensive data, leaving the nation partly blind to the emergence of virus variants. It also hinders the ability to assess virus contagiousness and severity or track the success of community and medical interventions. Crucial inputs are often unreliable or unavailable. Among the most important are the vaccination status of hospitalized patients as well as those who died. When vaccination status is recorded, it needs to be more than a binary choice between yes and no since there are multiple shots against Covid and other illnesses. Race, ethnicity, geographic, and other socio-demographic factors also need to be reported.

Much of the needed data are scattered across different and often incompatible computer systems. For example, Covid testing information is collected and reported by hospitals, physician practices, laboratories, independent testing sites, schools, employers, and nursing homes. Deaths may be reported by local or state authorities. Vaccinations are delivered and tracked by companies like CVS, health insurers, cities, and state health departments. Wastewater and genomic surveillance are reported by academic institutions and private firms, among others. Some of these data are shared with federal authorities. Some are not. Tracking it all down is all but impossible.

Getting information about how vaccine recipients fare after getting their shots is another crucial but nearly impossible task. In some cases, there are legal or privacy barriers to linking different types of data. In other cases, the systems just can’t handle it. Getting systems to talk to each other is a vital but expensive and time-consuming task since doing so is the only way the government can get the kind of information needed for an effective pandemic response.

Certain kinds of crucial information are not actively collected at all. Rather, the country relies on random, often incomplete reports and guesswork. Bad reactions or outcomes to drugs, vaccines and medical devices fit this category. Last year’s pause in the administration of the Johnson & Johnson vaccine to evaluate the frequency and severity of rare adverse reactions is just one example. Having a real-time, comprehensive data hub tracking and analyzing adverse reactions to vaccines would have either uncovered the problem far earlier or given officials the confidence to avoid the pause altogether.

The Covid pandemic has highlighted data problems researchers and public health officials have been complaining about for decades. Finally, legislators have begun to realize these problems aren’t just academic concerns but are threats to national security. Pandemic lockdowns cost trillions in economic activity, and each day’s delay in lifting them costs billions.
economical activity, and each day’s delay in lifting them costs billions. But the health information needed to guide such decisions in the United States is all but impossible to get. So, policy makers have been forced to rely on information from Great Britain, Israel and other countries where the data are better and more timely. This is unacceptable. American leaders need to know how an infectious agent is affecting Americans, who are unique in crucial socio-demographic ways. And they need to understand how it is traveling through New York and Dallas, not London or Tel Aviv. Why is the United States relying on other countries to fight its battles against deadly threats?

Why is the United States relying on other countries to fight its battles against deadly threats?

Establishing Centralized Public Health Data Platform(s)

Ultimately, the federal government must strive to build and operate a secure, standardized, and real-time national data platform for SARS-CoV-2 and other health threats. The United States already has data repositories for a considerable amount of public health data, but it is not standardized, coordinated, complete or timely. The closest approximation is the National Syndromic Surveillance Program (NSSP), a collaborative effort between the CDC and local public health stakeholders that collects, analyzes, and reports on patient encounters mostly in emergency rooms. The platform currently collects data from more than 6,000 health facilities across 49 states, receives data within 24 hours of initial encounters, and covers more than 70% of the country’s emergency departments. This data is then used to identify and inform responses to potential public health threats. The NSSP should either be expanded to collect and report on a broader set of standardized public health data, or a similar platform should be created to do so.

Standardizing Data Inputs and Quality

The reporting of some items is standardized while other categories are not, making for an often confusing mess. If everything were standardized, officials wouldn’t have to perform the time-consuming and arduous task of cleaning data. A classic example of this are death certificates, which would be a powerful tool to track serious health threats if the information on the certificates was consistent. It is not. For instance, the date on the document is sometimes the date of death, sometimes the date the death was reported and sometimes the date the certificate was created. Such differences matter. Agreement must be reached on how jurisdictions collect and report on:

- Case counts (type of tests, location of tests, deduplication of results, dating method)
- Positivity rate for testing (as above)
- Hospitalizations and hospital census (inpatient with information on ICU admission or not, and of vaccination status)
- Deaths
- Immunizations (clear identification and timing of 1st, 2nd, 3rd, Nth dose versus boosters, especially with heterologous vaccinations)
- Population immunity
- Animal reservoir testing (date of testing, population size, variant types)
- Wastewater sampling (date of sample, copies per milliliter)
- Genomic surveillance (variants tracked regionally and by percent of cases)
Death certificates would be a powerful tool to track serious health threats if the information on the certificates was consistent. It is not.

Notably, data collection should be prioritized, as government and other healthcare stakeholders have limited time and resources. FQHCs and other safety-net providers operate with especially constrained resources. Since the data they collect is crucial to understanding health disparities and designing policy responses, resources should be specifically granted to them and other organizations supporting underserved communities.

Some consideration should be given to establishing digital certificates that provide vaccination status and previous infection records to simplify tracking of breakthrough cases and reinfection frequency.

Linking Disparate Data Types and Sources

While data standardization would go a long way towards accelerating analysis and subsequent public health responses, many analyses are impossible or extremely difficult if data aren’t sufficiently linked. For instance, death reports and vaccination data might be standardized and separately available but figuring out which John Smith is which in two separate databases can be a bear. This is why even a seemingly simple analysis comparing deaths amongst the vaccinated and unvaccinated is often fraught.

Much of this isolation results from a simple lack of funding. Most recently, researchers have called for upgrading the country’s genomic surveillance systems to incorporate wastewater surveillance and link data to downstream clinical outcomes. While this would likely prove extremely helpful in early detection of variants, the cost would be substantial. At scale, the CDC’s Data Modernization strategy has not been fully funded but would take meaningful steps towards linking these data types and sources.

Similarly, the Office of the National Coordinator for Health Information Technology (ONC) has a significant role to play in collecting, linking, and storing various data sources but does not have sufficient resources to comprehensively do so. Public-private collaboration may be needed to encourage private data collectors to standardize and ease linkages across systems.

Notably, some linkages are difficult or currently impossible due to legal and regulatory barriers, many of them reflecting privacy concerns. For example, linking human genomic data to vaccination status and clinical outcomes runs into HIPAA compliance challenges, as genomic data is considered private and is theoretically impossible to de-identify since everyone’s DNA is unique. These privacy concerns have some merit, as an individual’s genomic data provides significant information on their health and could be used in a discriminatory manner. An independent advisory council composed of epidemiologists, data scientists, and public health decision-makers should be established to clarify and reevaluate regulations related to genomic data linkages and other data types where legal barriers impede public health response.
Standardizing Timely Reporting Cadences

Inconsistent schedules for reporting data can add to confusion. The government must rationalize reporting timelines. Results and data must be rapidly conveyed to both patients and public health officials, to promote both good patient care and effective pandemic response. Any entity that collects critical data must be able to report it to a centralized repository within 24 hours, as has been achieved by the NSSP.

Accelerating Insight Generation and Provide Public Access to Data

Investments in automated data analysis and insight generation are needed to realize the full benefits of any centralized repository. Automated analysis should prioritize real-time evaluation of incoming public health data, and algorithms should be developed to translate these analyses into actionable insights and practical public health guidance.

Critically, the federal government needs to be able to host data in ways that can be used by leaders, researchers, journalists, and the general public. Repositories must have access points for secure downloading and uploading of data, as well as easy-to-use visualizations, graphics, and analytical tools. And it must be appropriately staffed, including round-the-clock technical support for users.

Fifty states, hundreds of municipalities, and even universities have established best practices for public reporting that can be leveraged. As an example, Arizona has a rich state-level dashboard.82 Israel also has a transparent and consistent means of reporting most vital health measures quickly and publicly.83 Modeling a public reporting hub on the Covid Tracking Project would provide a strong starting point, given its strengths as a centralized one-stop shop for all Covid-related metrics, clear explanation of standards, graphical representations of data for use by journalists and the public, and open access to data.84

Since journalists, researchers, and the general public will all access the site, explanations about the data and reports available must clearly delineate what is modeled, what is adjusted, and what is raw, unadjusted data.

---

Designing Infrastructure to Promote Health Equity

America’s health data infrastructure can and should be redesigned to promote health equity. Sufficient standardized collection of socio-demographic data, including race, ethnicity, and sex linked to key medical data is necessary to evaluate disparities in incidence, treatment, and outcomes.

As public health experts’ usage of algorithms, machine learning, and artificial intelligence accelerates, efforts must be made to reduce programming biases. In one of many such examples, an algorithm attempting to predict patient health using health care costs as a proxy mistook Black patients’ lower spending on health care as an indicator of better health, suggesting less need for follow-up care. As public health experts’ usage of algorithms, machine learning, and artificial intelligence accelerates, efforts must be made to reduce programming biases. In one of many such examples, an algorithm attempting to predict patient health using health care costs as a proxy mistook Black patients’ lower spending on health care as an indicator of better health, suggesting less need for follow-up care.85 A similar study found that algorithmic genetic risk predictions were less accurate for non-White populations as the dataset used to inform predictions consisted of data primarily from White study participants.86 These inaccurate and biased algorithms are likely to exacerbate health disparities. Going forward, algorithms must be built on data that is representative of the populations being served, and public health guidance must be tailored to communities as well as specific ethnic and racial groups.

Notably, the January 25th draft of the bipartisan PREVENT Pandemics Act would authorize a new program to evaluate and identify best practices when collecting demographic information to support public health responses. This provision has the potential to play a meaningful role in improving linkages across demographic and health outcomes data, and it could make meaningful strides towards improving socio-demographic data collection that supports advances in health equity.

---

Acknowledging and Overcoming Implementation Challenges

None of this will be easy. Federal agencies and local actors have made significant strides in overcoming some challenges but have long faced barriers that are worth special mention.

None of this will be easy.

Money is the primary one. There has not been sufficient or dependable funding at either the federal or state levels to truly modernize America’s health data infrastructure. Even in instances when the federal government made necessary investments, infrastructure gaps and outdated systems at the local level impeded standardized and timely data collection. Most recently, the National Center for Health Statistics has taken the lead on attempting to modernize the country’s health data infrastructure but has struggled to overcome insufficient investments and engagement at the state and local levels. Providing performance-linked federal funding for state modernization efforts may accelerate state innovation and promote cooperation.

Another useful case study to consider is the National Vital Statistics System (NVSS), which collects direct and indirect mortality data and has received significant attention during the Covid pandemic. The limitations of the NVSS are well known—data collection is decentralized, there is a shortage of forensic pathologists and other critical staff, states are not required to report data to federal authorities, the CDC only partially funds data acquisition and analysis, and automation is limited. All of these problems have clear solutions. Among them are mandating state reporting, charging non-government users of the data like insurers to establish new revenue streams that fund collection and reporting efforts, bolstering forensic pathologists’ salaries and establishing loan forgiveness programs, and investing in sophisticated artificial intelligence tools that can minimize manual efforts. Funding limitations have long hindered progress in implementing these solutions. Given the importance of the NVSS’s data collection and reporting efforts in the midst of the pandemic, the funds should finally be made available.87

Another problem is that the size and complexity of public health data have ballooned and will continue to do so. By 2030, there will be 25 petabytes of genomics data.88 The country will need to incentivize and deploy substantial advances in machine learning and artificial intelligence to analyze data at this scale.

Lastly, the country’s health data infrastructure is complicated by numerous legal barriers. While some of these restrictions reflect legitimate privacy concerns, an independent advisory board should conduct a comprehensive reevaluation of the hodgepodge of laws and regulations that have been developed over the last 50 years.

---

Health Data Infrastructure Strategic Goals

1. **Empower and fund the CDC to rapidly develop standardized, national, real-time, comprehensive, and secure data platform(s) to monitor respiratory viruses and illnesses.**
   
   a. Require and provide sufficient funds for the CDC to evaluate existing national health data repositories, to understand current gaps and opportunities for consolidation.
   
   b. Require the CDC to augment existing health data repositories or create new secure data platforms that can collect comprehensive, standardized, and anonymized data from states, localities, and health care providers on major respiratory viruses and illnesses.
   
   c. Eliminate overlapping reporting requirements between the CDC and other federal agencies.

2. **Direct HHS to establish consistent national data standards by identifying critical metrics, defining them clearly, and establishing collection, linkage, and reporting requirements.**
   
   a. Evaluate and select a limited set of real-time and near real-time metrics that provide meaningful information on respiratory virus spread and severity, prioritizing those that are straightforward to collect, report, and analyze, as well as those that meaningfully inform public health decision-making. These metrics likely include hospitalizations and positivity rates, cross checked against vaccination or prior infection status and comorbidities.
   
   b. Invest in automation of data collection and reporting wherever possible, to minimize impact on delivery system.
   
   c. Establish standardized reporting pathways and protocols, including collection and reporting cadence, data quality standards, and triage approaches.
   
   d. Fund the CDC and make resources available to the private sector to collaboratively develop processes and protocols to link anonymized data across disparate sources into a single centralized destination and identify required and/or recommended data linkages. (e.g., test results, outcomes, vaccination status, age, sex, race, ethnicity, and other socio-demographic information).
   
   e. Establish an independent advisory council to reevaluate legal and regulatory barriers to linking disparate data types where these barriers impede public health response, reporting to the CDC within 90 days.

3. **Direct HHS to financially incentivize or require real-time reporting from states, localities, health providers and at-home test-takers, to include secure and de-identified test results, vaccination status, vaccine breakthrough and re-infection status, age, sex, ethnicity, race, job, workplace, and other essential socio-demographic information.**
   
   a. Fund and incentivize states, localities, health providers, and other stakeholders to establish automated data collection and reporting systems, as well as maintain, routinely update, and upgrade these systems.
   
   b. Require reporting on above data types from states, localities, and health providers, across the public and private sectors.
   
   c. Create a pathway to incentivize automated individual reporting of verified at-home tests. One way to do this might be to offer a dollar for each home test result returned.
   
   d. Launch nationally funded, duration-limited regional data hubs to provide transitional support to states, health providers, and other stakeholders, that are responsible for the implementation of reporting systems and standards.
   
   e. Create direct linkages from the CDC to state and local...
public health officials, as well as health providers and other public and private stakeholders, to establish bi-directional communication channels and promote collaboration.

4. **Direct and fund the CDC to accelerate insight generation and provide open access to data.**

   a. Invest in machine learning, artificial intelligence, and algorithms that automate analysis and speed development of actionable insights.
   b. Establish a centralized real-time reporting dashboard, for Covid and other respiratory viruses, with easy-to-read key metrics including hospitalizations, deaths, positivity rates, vaccine breakthroughs / re-infections, and long Covid cases, and link metrics to designated public health actions triggered by escalations in disease spread and severity.
   c. Produce regular actionable reports for the public, municipalities, states, and national decision-makers to inform public health actions.
   d. Establish a real-time open data hub, with both pre-built analytical tools and de-identified raw data files, that provides access to reporters, researchers, and other interested parties.
   e. Establish a searchable, real-time research hub for pre-print and peer-reviewed analyses managed by the CDC (or a new agency), with staff escalating critical findings to agency leadership.

5. **Direct HHS and the CDC to design health infrastructure to promote health equity and mitigate racial, ethnic, occupational and gender disparities.**

   a. Collect sufficient standardized socio-demographic data including race, ethnicity, and sex, linked to key medical data, to better evaluate disparities in incidence, treatment, and outcomes.
   b. Base algorithm design on datasets that are representative of target populations, to protect against bias and discrimination.