What Data Should Be Included in a Modern Public Health Data System

Joie D. Acosta,1 Anita Chandra,1,* Douglas Yeung,2 Christopher Nelson,2 Nabeel Qureshi,3 Tara Blagg,3 and Laurie T. Martin1

Abstract
The public is inundated with data, both in where data are ubiquitously collected and in how organizations are using data to drive public sector and commercial decisions. The public health data system is no exception to this flood of data, both in growing data volume and variety. However, what are collected and analyzed about the health status of the nation, how particular data and measures are prioritized for parsimony, and how those data provide a signal for where to invest to address health inequities are in dire need of a reboot. As with other articles in this supplement, this article builds from a literature review, an environmental scan, and deliberations from the National Commission to Transform Public Health Data Systems. The article summarizes what data should be included and identifies where the technology and data sectors can contribute to fill current gaps to measure equity, positive health, and well-being.

Keywords: equity; public health; measures; well-being

Introduction
It can be argued that there is a significant volume of data on health and well-being, but that these data are not providing the information needed to fully advance effective and equitable public health action today. The role of equity is central to a modern public health data system, and the recognition that the current data system falls short in that objective is clear. However, there are other interrelated challenges to identifying and tracking populations affected by multiple stresses, such as COVID-19 and chronic disease. There is an opportunity to reimagine what matters for health in the United States; how the nation uses data as a tool to aid action on inequity; and how the United States uses public health data to capture concepts, such as systemic health injustice, well-being, and resilience.

In this article, we describe the types and content of data needed in this reimagined public health data system. Advancing the culture of health, one that is guided by the cultural and social drivers of health, requires a different approach to public health data collection and use, including how to move beyond a focus on illness and capture concepts such as positive health and well-being.1 To date, public health data have not always led to appropriately targeted public health action, in large part, because there is no integrated or coordinated public health data system to support sense-making (described in the prior article) and data-informed decision-making. As with the move to an equity-centered public health data system, the opportunities to pool new and different types of data provide a window for many sectors, outside formal public health, to engage in new ways.

With access to a greater variety of data and potentially faster use of data, the data science and technology sector has an important role to play.

1RAND Corporation, Arlington, Virginia, USA.
2RAND Corporation, Santa Monica, California, USA.
3Pardee RAND Graduate School, Santa Monica, California, USA.

*Address correspondence to: Anita Chandra, RAND Corporation, 1200 South Hayes Street, Arlington, VA 22202, USA, E-mail: chandra@rand.org

© RAND Corporation 2022. Published by Mary Ann Liebert, Inc. This Open Access article is distributed under the terms of the Creative Commons License [CC-BY] (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
In 2020, the Robert Wood Johnson Foundation formed the National Commission to Transform Public Health Data Systems to review significant challenges to the current public health data system, and “provide recommendations to policymakers, health care organizations and institutions, service providers, and philanthropy” on potential solutions to overcome these challenges. In support of this effort, RAND conducted a supporting analysis that included an environmental scan to identify key issues, points of consideration, trade-offs, and tensions, and current activities related to public health data, data systems, and data modernization efforts. This effort included a targeted scan of published research articles and reports, reviews of websites and working documents describing coordinated activities (e.g., data interoperability), and recent initiatives. Additional searches included the use of “big data” in public health, data privacy, and ethics of public health data collection. Although the team primarily focused on public health data, it also identified seminal articles and reports from other sectors or disciplines whose findings could apply to public health data systems.

RAND simultaneously conducted semistructured interviews with 112 experts and thought leaders on the main topics before the Commission. Individuals represented diverse sectors, including public health and health care, technology and data science, research and policy, journalism, and law. The interviews also included experts in data, data use, equity, community engagement, and research translation who work outside the traditional health sector. The project was reviewed and approved by the RAND Human Subjects Protection Committee.

In this article, we highlight relevant findings from this supporting analysis and then implications for the data science and technology sectors, with consideration of recommendations that emerged from the final Commission report.

**Findings**

Four major themes emerged from the supporting analysis regarding the approach to collecting and reporting public health data; the content of public health data; the volume and coordination of data across sectors, particularly unstructured data; and the level of precision, granularity, and timeliness of data currently reflected in public health data sources. These themes were further affirmed in the Commission recommendations related to ensuring public health data include information on structural racism and other drivers of inequity.

In the following section, we briefly describe these four themes.

The modern public health data system is reactive rather than proactive and covers a wide array of measures rather than a smaller prioritized set. The main function of public health data has remained consistent over the years, primarily focused on population health surveillance, one of the 10 essential public health services. However, data are used to support other public health services, including the selection of public health actions, communication with the public, and the building of partnerships to improve health. Despite these varied uses of public health data, the content of the data has been augmented in ways that are more often reactive in monitoring health conditions. This often includes the disproportionate use of lagging indicators (meaning data on performance today or what has occurred), rather than leading indicators that portend future health needs.

Data systems have been enhanced as the nation has learned more about key factors that influence health.

Healthy People, the national strategic management plan that guides health promotion and disease prevention, demonstrates the steady expansion of public health’s primary areas of focus over the past 40 years. This focus has expanded from a singular emphasis on mortality to now include aspects of quality of life, health disparities and health equity, and social and physical environments. Yet, despite the expansion in these focus areas, the data and indicators that are often a part of public health data emphasize infectious disease, chronic conditions, mortality, and risk factor exposures, with less attention to other factors that influence health over generations (e.g., trauma), as well as measures of positive health and well-being and systemic inequities.

In addition to the need for data that can support a proactive response is the need for narrowing the data that are used to facilitate that response. Modernizing the public health data system requires prioritizing a smaller

---

4 Ten essential public health services: (1) Assess and monitor population health status, factors that influence health and community needs and assets. (2) Investigate, diagnose, and address health problems and hazards affecting the populations. (3) Communicate effectively to inform and educate people about health, factors that influence it, and how to improve it. (4) Strengthen, support, and mobilize communities and partnerships to improve health. (5) Create, champion, and implement policies, plans, and laws that impact health. (6) Utilize legal and regulatory actions designed to improve and protect the public’s health. (7) Assure an effective system that enables equitable access to the individual services and care they need to be healthy. 8. Build and support a diverse and skilled public health workforce. (9) Improve and innovate public health functions through ongoing evaluation, research, and continuous quality improvement. (10) Build and maintain a strong organizational infrastructure for public health.
set of core national measures to ensure the parsimony needed to proactively make social change on a few key public health efforts, rather than continue to spread the primary focus of public health thin across many areas.3 If addressing systemic health injustice becomes the priority, then the primary areas of focus and the associated measures and indicators should emphasize upstream drivers of inequity before those inequities become disparities.

However, the current use of the public health data system in the United States does not proportionately weight measures toward any clearly operationalized priority, whether positive health, addressing inequity, or other forward-leading public health action. As such, the current public health data system does not fully communicate a common set of values through an alignment and parsimony of national measures. The content of public health data misses important information about structural inequity and positive health and well-being

**The content of public health data misses important information about structural inequity and positive health and well-being**

Even with increasing amounts of health data being collected, public health data are limited in their ability to inform decision-making because data on many upstream factors that contribute to health are not regularly or consistently collected. A review of the Healthy People 2030 Leading Health Indicators by social determinants of health suggests continued gaps in current public health data, particularly related to economic stability, neighborhood and the built environment, education, and community and social context.6 A broader review of national public health data sources conducted for this supporting analysis revealed similar gaps.3

There are several U.S. efforts underway that are working to include broader conceptualizations of well-being in surveys. This includes the Gallup survey, which captures subjective well-being (e.g., optimism, hope, resilience) and happiness7; Measure of America, which includes some sentinel measures of well-being and equity8; and Well-Being in the Nation, which includes measures of well-being and has advanced the consistent use of the Cantril ladder, a measure of life evaluation and expectation linked to other measures of mental health and well-being in a population.9 Healthy People 2030 has begun to integrate subjective well-being and to push those measures in national surveys, such as the Behavior Risk Factor Surveillance System.

The Centers for Disease Control and Prevention has integrated well-being concepts for some time, yet currently those are mostly linked to Health-Related Quality of Life.10 However, areas such as lifelong learning and prosocial health behaviors, useful constructs for understanding drivers of health outcomes, are noticeably absent even from these efforts to advance well-being assessment. Although there has been advancement in the use of surveys to capture subjective well-being and understand the drivers of well-being, there are opportunities to use more information coming from the private sector to fill these gaps and consistently link information on sentiment and community needs to inform public health action. To date, these efforts are not standardized.11

**Handling the volume of new public health data requires greater support and coordination, particularly for unstructured data**

Unstructured data refers to nonstandardized data from a variety of sources (e.g., social media, clouds, sensors) and can include text, images, audio, video, blogs, websites, and so forth.12–16 Unstructured data are growing exponentially, with more than 2.5 quintillion bytes of unstructured data generated every day. Unstructured data are expected to continue to grow over time, with an estimated 65 percent annual growth rate, and to comprise ~95 percent of global data.14 These unstructured data can provide critical insights into public health; however, these data have not yet been effectively integrated into public health data systems. For example, clinical notes and other free text from electronic health records have been analyzed for insights into patient care and overprescribing and use of certain medications.17

Text from social media posts has been analyzed to look for a positive or a negative sentiment providing insight into the psychological well-being of a population, as well as self-reported symptoms of a disease to help estimate disease outbreak.18,19 There is emerging research about the use of images posted on social media to track drinking, smoking, and obesity-related behaviors and the use of wearable sensors for remote patient monitoring and virtual health assessments.20,21 The huge volume and complexity of existing and unstructured data make it a tedious task to extract useful information from different types of data, and its use is fraught with concerns about ethical trade-offs. More data are not useful unless they can be used for proactive decisions.

While past research has demonstrated how unstructured data can inform public health decision-making, these investigations have raised concerns about how
to maintain privacy, especially if artificial intelligence is used to generate or distill information that could be considered personally identifiable medical information. Data ownership is another ethical challenge identified in using unstructured data and which will require strong governance policies.

Precision, granularity, and timeliness of public health data continue to challenge the ability to make responsive decisions. The types of public health data have adapted according to the changing public health data and informatic needs, moving from counts and trends to causal inferences to geospatial inferences over time. Public health practitioners often rely on federal data systems to create those inferences, but those data must be collected, cleaned, and deidentified, creating significant time lags often measured in years. This requires significant resources and can hinder public health departments in making timely decisions because they are constantly relying on survey data that may be lagging by a year or more.

Making key decisions with outdated information can negatively impact public health, particularly if populations are dealing with a public health emergency. For example, the lack of timely, federal data made it more difficult to inform a pandemic containment strategy and shed new light on the limitations of our current public health data system for supporting real-time decision-making. While countries such as South Korea, Singapore, and New Zealand were disseminating public health data in near real-time, the U.S. data lagged and were often lacking critical information, such as race and ethnicity.

Implications
The stakeholder analysis, review of existing data systems, and Commission deliberations identified key areas in which the data science and technology sector can be particularly useful. For example, one of the Commission’s recommendations was focused on shifting the national health narrative to one that is equity-based and positively oriented. As noted in the findings described earlier, there is a large gap in the data available (both in time and quality) on positive health, well-being, and the drivers of structural inequity, some of which could be informed by the data science and technology sector specifically. Furthermore, data science and technology experts often have greater understanding of how to handle the dimensions of “big data” in terms of data structure, precision, and data volume. We further describe those two key implications for this sector with respect to the content of public health data.

The technology sector has information that can support public health action regarding positive health, well-being, and structural inequity. Many companies are examining the data they collect and own and are seeking to repurpose those data for public good. For instance, person-level, transactional data exist on almost every aspect of our lives and are available to private industry. Technologies, such as global positioning system and accelerometry data from wearable devices, provide insight into bodily movements and location. Purchasing and travel activity provide insight on personal values, socioeconomic status, and healthy behaviors. Online habits, such as web searches or browsing histories, can illuminate what people want to know about health, as well as offer sensitive markers of health status. Because of network effects or first mover effects (i.e., the benefits accrued from being the first to market), much of the data for a given technology type are monopolized by a single dominant company in that domain.

Technology companies that seek to leverage their own products and holdings for social good are largely doing it on their own—that is, looking inward to their own data and tools and focusing on their own capabilities. One way in which some companies (e.g., Twitter) have tried to do this is by making subsets of their data openly available for researchers or government use, under specific data use or nondisclosure agreements. However, these data tend to be scrubbed of personal identifying information and limited in scope or by region, which may be necessary for privacy and data security, but can limit their usefulness for public health action. Currently, public health does not have a consistent, established relationship with technology companies to obtain or leverage the data or innovative analysis of unstructured and “big data” that many of these companies are undertaking. However, a few notable exceptions provide promising models for such collaborations.

The data science and technology sector can offer approaches for balancing precision and granularity with real-time response. Existing public health data challenges related to the speed and precision trade-off may be overcome by
learning from and partnering with technology companies, which have resources and data capabilities that, in many cases, far exceed the typical public health sector. These companies have data that are already continually collected and analyzed. Using massive amounts of data in combination with sophisticated analytic techniques, companies can create accurate profiles of individuals or population segments based on numerous characteristics. These capabilities can be used to rapidly identify behavior changes, to answer emerging questions, or to spot shifts in sentiment. Social media posts or web searches provide real-time granular information that can reveal mental states such as depression or anxiety, or whether someone has searched for information about how to hurt themselves or others. 

Many products that collect data are already in our households, which could help lessen a reliance on traditional data collection procedures that are more laborious and take longer to implement. One impact of this could be to increase the representation in public health data of an otherwise-hidden population. 

Attention should still be paid to individuals who may not be online due to the lack of broadband in some regions, limited access to free public internet, inconsistent access to the internet due to economic instability, and varying degrees of digital literacy. It will be critical to acknowledge and address this digital divide to not create further inequity among populations that bear significant public health burdens but are not reflected in online data.

**Conclusion**

The content of current public health data for proactive public health action and equity-centeredness is insufficient. The data science and technology sector has an important, although still largely untapped, role in supporting how public health captures more temporally relevant and actionable data. For the relationship between formal public health and the technology sector to leverage the new window for public health data system modernization, the two sectors will need to have open and consistent dialogue about the content of available data, its utility for certain types of public health action, and the governance required to support ethical and equitable application.

**Author Disclosure Statement**

No competing financial interests exist.

**Funding Information**

This article was supported under a grant from the Robert Wood Johnson Foundation. The views expressed are solely the authors.

**References**

1. Hanlon P, Carlisle S, Hannah M, et al. Making the case for a “fifth wave” in public health. Public Health 2011;125(1):30–36; doi: 10.1016/j.puhe.2010.09.004.
2. National Commission to Transform Public Health Data Systems. Charting a Course for an Equity-Centered Data System: Recommendations from the National Commission to Transform Public Health Data Systems. Robert Wood Johnson Foundation: Princeton, NJ; 2021.
3. Acosta J, Chandra A, Martin L, et al. Transforming public health data systems, What? The data in the modern public health data system. Robert Wood Johnson Foundation, National Commission to Transform Public Health Data Systems: Princeton, NJ; 2021. Available from: https://www.rwjf.org/en/library/research/2021/10/charting-a-course-for-an-equity-centered-data-system.html [Last accessed: April 1, 2022].
4. Centers for Disease Control and Prevention. 10 Essential Public Health Services. Atlanta, GA; 2021. Available from: https://www.cdc.gov/publichealthgateway/publichealthservices/essentialhealthservices.html [Last accessed: 2022].
5. Developing Healthy People 2020. U.S. Department of Health & Human Services. Appendix 7, Building on Past Challenges of the Healthy People Initiative. Washington, DC; 2020; pp. 63–67. Available from: https://www.healthypeople.gov/sites/default/files/Phase1_0.pdf [Last accessed: 2020].
6. Ochiai E, Kigenyi T, Sondik E, et al. Healthy people 2030 leading health indicators and overall health and well-being measures: Opportunities to assess and improve the health and well-being of the nation. J Public Health Manag Pract 2021;27(1):S235–S241; doi: 10.1097/PHH.0000000000001424.
7. Clifton J. Gallup. The State of World Happiness in 2019. 2019. Available from: https://news.gallup.com/opinion/gallup/247940/state-world-happiness-2019.aspx [Last accessed: 2021].
8. Lewis K, Burd-Sharps S. Measure of America of the Social Science Research Council. The State of America’s Well-Being. Brooklyn, NY; 2013. Available from: https://measureofamerica.org/blog/2013/03/the-state-of-americas-well-being/ [Last accessed: 2020].
9. Well Being in the Nation Network. Well Being in the Nation (WIN) Measures. Hyattsville, MD; Available from: https://www.winmeasures.org/statistics/winmeasures [Last accessed: 2022].
10. Centers for Disease Control and Prevention. Health-Related Quality of Life (HRQOL) Well-Being Concepts. Atlanta, GA; 2018. Available from: https://www.cdc.gov/hrqol/wellbeing.htm [Last accessed: 2021].
11. Luhman M. Using Big Data to study subjective well-being. Curr Opin Psychol 2017;18:28–33; doi: 10.1016/j.copsyc.2017.07.006.
12. Lomotey RK, Deters R. Topics and terms mining in unstructured data stores. 2013 IEEE 16th International Conference on Computational Science and Engineering, Sydney, NSW, 2013; pp. 854–861; doi: 10.1109/CSE.2013.129.
13. Lomotey RK, Deters R. RSenter: Terms mining tool from unstructured data sources. Int J Busin Process Integrat Manag 2013;9(4):298–311; doi: 10.1504/IJBPIM.2013.059136.
14. Lomotey RK, Jamal S, Deters R, (eds). SOPHRA: A mobile web services hosting infrastructure in mHealth. 2012 First International Conference on Mobile Services, Honolulu, HI, 2012; pp. 88–95; doi: 10.1109/Mob-Serv.2012.14.
15. Scheffer T, Decomain C, Wrobel S, editors. Mining the Web with active hidden Markov models. Proceedings 2001 IEEE International Conference on Data Mining, San Jose, CA, 2001; pp. 645–646; doi: 10.1109/ICDM.2001.989591.
16. Wang Y, Kung LA, Byrd TA. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Change 2018;126:3–13; doi: 10.1016/j.techfore.2015.12.019.
17. Luo Y, Thompson WK, Herr TM, et al. Natural language processing for EHR-based pharmacovigilance: A structured review. Drug Safety 2017; 40(11):1075–1089; doi: 10.1007/s40264-017-0558-6.
18. Fung ICH, Tse ZTH, W. FK. The use of social media in public health surveillance. Western Pac Surveill Resp J 2015;6(2):3; doi: 10.3635/WPSAR.2015.6.1.019.
19. Wongkoblap A, Vadillo MA, Curcin V. Researching mental health disorders in the era of social media: Systematic review. J Med Internet Res 2017; 19(6):e228; doi: 10.2196/jmir.7215.
20. Garimella VRK, Alfayad A, Weber I, (eds). Social media image analysis for public health. 2016 CHI Conference on Human Factors in Computing Systems, 2016.
21. Seshadri DR, Davies EV, Harlow ER, et al. Wearable sensors for COVID-19: A call to action to harness our digital infrastructure for remote patient monitoring and virtual assessments. Front Digital Health 2020;2:8; doi: 10.3389/fdgth.2020.00008.
22. Adnan A, Akbar R. An analytical study of information extraction from unstructured and multidimensional big data. J Big Data 2019;6:91; doi: 10.1186/s40537-019-0254-8.
23. Wang YC, DeSalvo K. Timely, granular, and actionable: Informatics in the public health 3.0 era. Am J Public Health 2018;108(7):930–934; doi: 10.2105/AJPH.2018.304406.
24. Chan JL, Purohit H. Challenges to transforming unconventional social media data into actionable knowledge for public health systems during disasters. Disast Med Public Health Prepared 2020;14(3):352–359; doi: 10.1017/dmp.2019.92.
25. Maxmen A. Why the United States is having a coronavirus data crisis. Nature 2020;585:13–14; doi: 10.1038/d41586-020-02478-z.
26. Robbins R. Tech companies are crucial players in the coronavirus response. Are they contributing what’s most needed? Stat News. 2020. Available from: https://www.statnews.com/2020/03/25/tech-companies-coronavirus-response/ [Last accessed: 2021].
27. Romm T, Dwoskin E, Timberg C. U.S. government, tech industry discussing ways to use smartphone location data to combat coronavirus. The Washington Post; 2020. Available from: https://www.washingtonpost .com/technology/2020/03/17/white-house-location-data-coronavirus/ [Last accessed: April 15, 2022].
28. Stephens-Davidowitz S. Dr. Google Will See You Now. The New York Times; 2013. Available from: https://www.nytimes.com/2013/08/11/ opinion/sunday/dr-google-will-see-you-now.html [Last accessed: April 15, 2022].
29. Stephens-Davidowitz S. Fifty States of Anxiety. New York Times. 2016. Available from: https://www.nytimes.com/2016/08/07/opinion/sunday/ fifty-states-of-anxiety.html [Last accessed: April 15, 2022].
30. Jahani E, Sundsøy P, Bjelland J, et al. Improving official statistics in emerging markets using machine learning and mobile phone data. EPJ Data Sci 2017;6(1):3; doi: 10.1140/epjds/s13688-017- 0099-3.
31. Leinen S. Mobile Lifestyles: The Homeless and Their Cell Phones [Doctoral Dissertation]. Regent University; 2017.
32. Shi L, Stevens GD. Vulnerable populations in the United States. John Wiley & Sons: Hoboken, NJ; 2021.

Cite this article as: Acosta JD, Chandra A, Yeung D, Nelson C, Qureshi N, Blagg T, Martin LT (2022) What data should be included in a modern public health data system. Big Data 10:S1, S9–S14, DOI: 10.1089/ big.2022.0205.