Translational Health Disparities Research in a Data-Rich World

Despite decades of research and intervention, significant health disparities persist.[1] The National Institute on Minority Health and Health Disparities has identified a set of multilevel, complex and dynamic determinants of health disparities.[2] Health disparity populations are defined as racial/ethnic minorities, persons with low socioeconomic status, underserved rural residents, and sexual and gender minorities.[3] However, to date, limitations in data sampling, data collection, and analysis techniques have prevented us from developing a better understanding of causes of disparities and how to translate evidence into effective interventions that could reduce disparities while improving health outcomes for all. Below we present five big data approaches that could be leveraged to enhance translational health disparities research. These approaches for exploiting big data could help operationalize opportunities and challenges concerning use of big data science to reduce health disparities.[4] Importantly, we emphasize the ethical considerations in the analysis and interpretation of results, and the caution needed in making broad generalizations from big data sources before knowing whether they are indeed representative of the populations under study.

Various types of big data, such as geospatial, electronic health record, sensor, and molecular "omic" are collected, largely independently, through a mix of government, nonprofit, academic, and commercial entities. Studies using each of these data types have shown promise for discovery, for elucidating more proximate causes of disease, and for suggesting approaches for improving health equity. Although each data type on its own has contributed to health disparities research, at least two major challenges remain. First is how to integrate diverse data types for in-depth data analysis. Significant investments will be required to explore integrative approaches to using big data for health disparities research. Second, many such data sources are convenience or volunteer samples. Thus, these data represent potentially biased samples of behavior, environment and health outcomes from poorly defined populations. Much more work is needed to understand and adjust for possible sampling biases.

FIVE APPROACHES FOR USING BIG DATA

Table 1 illustrates five big data approaches that can be leveraged to enhance translational health disparities research. These include linking structured data, harmonizing data elements, fostering citizen science, developing large comprehensive cohorts, and mining novel data to improve disparities surveillance. Each approach offers major opportunities to better document health disparities, to better identify their underlying causes, and to evaluate efforts to reduce disparities. Each approach also has major limitations, highlighting the need for multiple approaches for improving and adding to existing data, and more applied research on how to link novel data. The first limitation is that, except for census and federal health surveys, most big data for health disparities suffer from significant risks of selection bias. Projects touting assessment of disparities with new data resources should make efforts to understand how such biases may influence their results no matter how "big" the data may be. Indeed, the failures of Google Flu Trends, a notable effort to use data mining to better predict disease, caution against "big data hubris," the notion that big data can substitute (rather than supplement) traditional data collection.[5] Results from big data may be meaningful but are difficult to interpret without more context than a Google search can provide.[6] Similar cautions may well apply to citizen science efforts; citizen scientist volunteers might not accurately represent their communities. A second is that maintaining engagement and ensuring diversity are significant challenges for research and voluntary data collection. The "All of Us" cohort is an effort by the National Institutes of Health to address these challenges in a very large and lengthy health study. Finally, harmonizing data elements and identifying new measures of the contextual influences on disparities require considerable cooperation and research, and these methodological and consensus activities can be difficult to fund and sustain.

Novel big data could supplement federal data to document local disparities and disparities in small populations, reveal the causes of health disparities, and allow evaluation of programs and policies at multiple spatial scales. Each data type needs to be adequately documented with metadata describing the elements, sources, and assessment of quality. This documentation is critical

ABOUT THE AUTHORS

Nancy Breen and Xinzhi Zhang are with the National Institute on Minority Health and Health Disparities (NIMHD), National Institutes of Health (NIH), Bethesda, MD. James S. Jackson is with the Institute for Social Research, University of Michigan, Ann Arbor. Fred Wood is with the National Library of Medicine, Bethesda, MD. David W. S. Wong is with the Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA. Nancy Breen is also a Guest Editor for this supplement issue.

Correspondence should be sent to Nancy Breen, 6707 Democracy Blvd Suite 800 MSC 5465, Bethesda, MD 2892 (e-mail: breenn@mail.nih.gov). Reprints can be ordered by clicking on the "Reprints" link at www.ajph.org.

This editorial was accepted December 22, 2018.

Note. *The content is the responsibility of authors and does not necessarily represent the perspective of the US government.*

doi: 10.2105/AJPH.2019.304951

TABLE 1—Five Approaches to Enhancing Use of Data and Big Data for Health Disparities Research and Evaluation

Approach	Example	Exemplary Challenge	Resources
Linking structured data	SEER-Medicare	Limited to ages 65 y and older	https://healthcaredelivery.cancer.gov/seermedicare
Harmonizing data elements	Cancer Research Network	Limited to insured	https://crn.cancer.gov
Fostering citizen science	Our Voice project	Scaling up existing small projects	http://med.stanford.edu/ourvoice.html
Developing big longitudinal cohorts	NIH All of Us cohort	Sustaining engagement	https://allofus.nih.gov
Mining Internet and social media data to improve disparities surveillance	Google searches for disease and disparity keywords	Validation, big data hubris, algorithm dynamics and stability	https://gking.harvard.edu/files/gking/files/0314policyforumff.pdf https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0122963

Note. NIH = National Institutes of Health; SEER = Surveillance, Epidemiology, and End Results program.

for combining these data sources. A proposed standard to make data more available and interoperable is the FAIR (Findable, Accessible, Interoperable and Reusable) principles.[7] In addition to these core standards, the field of data science needs to address acquisition, engineering, curation, storage, analytics, visualization, dissemination, ethics; as well as legal, policy, and social impact. Each element represents a distinct challenge for the application of data science and big data resources to health disparities research and translation. Mechanisms to promote close collaboration between data scientists and health disparities experts are needed to maximize the utility and reuse potential of investments in data collection and health disparities research.

ETHICAL RESPONSIBILITIES, OTHER CHALLENGES

People experiencing health disparities, researchers, program and policy staff, and leaders addressing health policies present a spectrum of opinion about the value of big data approaches. These differing opinions range from lack of trust, to acceptance, to enthusiastic endorsement. Researchers must be cognizant and respectful of these differences because building trust is a key aspect of building a broad coalition, which is needed for successful generalizability. Moreover, researchers have the responsibility to ensure their research does not cause harm to either individuals or communities.

Biomedical ethics usually is concerned with potential harm to individuals. Health disparities research requires coupling many different types of data, potentially permitting identifying individuals, which increases the risk for personal harm. In addition, communities experiencing health disparities may find their entire community being stigmatized by research findings that emphasize or overstate negative features. Therefore, health disparities researchers must be mindful of social as well as individual ramifications of data and findings. As big data enter minority health and disparities research, a key ethical concern will be the need to ensure that results equally benefit all populations. Ethical dilemmas associated with who should have access to data and the impacts of interpretation need to be constantly considered. These ethical issues need to be addressed when data capacity is being built, and not after the fact. A further key challenge concerns how best to share complex big data and results with study participants.

Today, unprecedented opportunities exist to broaden inquiries into health disparities using a growing spectrum of diverse data sources. Mindful of the threat of "data hubris," we are optimistic about the power of new data and new linkages. With the appropriate workforce and tools, careful analysis and triangulation of big data could lead to greater knowledge and greater potential for effective interventions that could not be imagined earlier in reducing health disparities. AJPH

Nancy Breen, PhD
James S. Jackson, PhD
Frederick Wood, DBA, MBA
David W. S. Wong, PhD
Xinzhi Zhang, MD, PhD
The Data in Health Disparities Writing Team

CONTRIBUTORS

All authors and members of the writing team conceived the study and contributed to the analysis. N. Breen led the writing.

ACKNOWLEDGMENTS

The Data in Health Disparities Writing Team includes: David Berrigan, PhD, Division of Cancer Control and Population Sciences, National Cancer Institute; Philip E. Bourne, PhD, Data Science Institute and Department of Biomedical Engineering, University of Virginia; and Joshua C. Denny, MD, MS, Biomedical Informatics and Medicine, Vanderbilt University Medical Center.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to report.

REFERENCES

1. Agency for Research and Healthcare Quality. 2016 National Healthcare Quality and Disparities Report. 2018. Available at: https://www.ahrq.gov/research/findings/nhqrdr/nhqdr16/index.html. Accessed December 10, 2018.

2. Alvidrez J, Castille D, Laude-Sharp M, Rosario A, Tabor D. The NIMHD research framework for minority health and health disparities. *Am J Public Health*. 2018; 109(S1):S16–S20.

3. Pérez-Stable EJ, El-Toukhy S. Communicating with diverse patients: How patient and clinician factors affect disparities. *Patient Educ Couns*. 2018;101(12): 2186–2194.

4. Zhang X, Pérez-Stable EJ, Bourne PE, et al. Big data science: opportunities and challenges to address minority health and health disparities in the 21st century. *Ethn Dis*. 2017;27(2):95–106.

5. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. *Science*. 2014;343(6176):1203-1205.

6. Auerbach D. Big anecdata. *Slate*. May 6, 2015.

7. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.

S42 Editorial Breen et al. AJPH Supplement 1, 2019, Vol 109, No. S1