Foreword: Big Data and its Application in Health Disparities Research

Guest Editors: Eberechukwu Onukwugha, MS, PhD[1];
O. Kenrik Duru, MD, MS[2];
Emmanuel Peprah, PhD[3]

The articles presented in this special issue advance the conversation by describing the current efforts, findings and concerns related to Big Data and health disparities. They offer important recommendations and perspectives to consider when designing systems that can usefully leverage Big Data to reduce health disparities. We hope that ongoing Big Data efforts can build on these contributions to advance the conversation, address our embedded assumptions, and identify levers for action to reduce health care disparities. *Ethn Dis.* 2017;27(2):69-72; doi:10.18865/ed.27.2.69.

Keywords: Big Data, Health Disparities

[1]Pharmaceutical Health Services Research Department; University of Maryland School of Pharmacy
[2]Division of General Internal Medicine; Geffen School of Medicine, University of California Los Angeles
[3]Center for Translation Research and Implementation Science (CTRIS); National Heart, Lung, and Blood Institute; National Institutes of Health

Address correspondence to Eberechukwu Onukwugha, MS, PhD; Pharmaceutical Health Services Research Department, University of Maryland School of Pharmacy, 220 Arch Street, Baltimore, MD 21201; 410.706.8981; eonukwug@rx.umaryland.edu

We developed this themed issue out of a desire to understand the relevant issues that arise when utilizing Big Data to document, explain and address health care disparities. The term, Big Data, has been coined to define the large and complex datasets that are characterized by volume (large number of records), variety (multiple sources), velocity (frequent, and not necessarily uniform, updates) and value (unique and powerful insights). We sought manuscripts from diverse settings including academia, health care systems, contract research organizations, and government. The contributed manuscripts offer an important, baseline snapshot of current perspectives on Big Data and health disparities. Specifically, the articles in this special section provide: 1) an overview of recent efforts to improve workforce diversity and training through the National Institutes of Health Big Data to Knowledge (BD2K) diversity initiatives; 2) a clinical trial perspective on whether Big Data will increase or decrease disparities within the health care system; 3) results from an investigation of how differential linkage by race can affect health care disparities research using the linked data; 4) results from a study of Hispanic residential isolation and the relationship with Attention Deficit Hyperactivity Disorder health service utilization; 5) a scoping review of opportunities for Big Data science to address health disparities as well as some challenges.

Although we are familiar with Big Data from a modern viewpoint, the problem of analyzing large amounts of data to understand the content of said data, draw reasonable conclusions, and make meaningful inferences is an age-old problem requiring the development of novel tools. Big Data problems existed even before the invention of the 20th century computer. For example, it was predicted that it would take more than eight years to hand-tabulate all the data collected in the 1880 US census, and it would take more than 10 years to analyze data from the 1890 census.[1] The US Census Bureau needed a major technological advance to analyze and produce accurate data to inform policy development at the government level. To accomplish this task, Herman Hollerith automated the process and developed a storing and information processing system in the 1880s.[2]

Data standardization, computation analysis, methodology, statistical approaches, and data storage all represented challenges that Hollerith had to solve in addressing his Big Data problem. We have similar challenges in the modern era. As an example, the development of genome-wide association studies (GWAS) required the creation of novel analytical pipelines that combined statistical analysis with computing power.[3] These GWAS analytical pipelines have had limited integration power because of the utilization of linear univariate methodology. Higher order interaction studies on large GWAS utilizing multivariate analysis are needed but currently computationally intensive and thus limited by the availability of hardware.[4] Moreover, characterization of pleiotropic effects based on integration of various Big Data types will be critical to identifying molecular pathways that underlie disease and health.[5] Using Hollerith's example, we note that these challenges provide the opportunity to develop solutions that can address the vexing problems posed by Big Data in the modern era.

Currently, 87% of physicians and 80% of non-Federal acute hospitals in the United States use single-system electronic health records (EHRs), in addition to the development of aggregated Clinical Data Repositories (CDRs) and other massive databases.[6] It is possible that novel uses of Big Data will finally reduce or eliminate stubbornly persistent racial/ethnic and socioeconomic disparities in health outcomes. The articles in this issue discuss the current data, privacy, and larger societal environments and illustrate how we can make progress toward this goal. Canner et al describe the ongoing efforts to increase workforce diversity within the field of biomedical data science. Creative projects that combine preexisting public databases, as illustrated by Miller et al and Pennap et al, can provide important insights in the delivery and outcomes of care. As noted by Heller & Seltzer, ignoring best practices for the use and interpretation of either clinical trial or observational data can limit our ability to identify and address health disparities. Most EHRs are currently constructed to maximize billing revenue; increasing the capacity of these datasets to collect and merge clinically relevant data elements will be an important step toward building the infrastructure needed to tackle health disparities.

Datasets built on personalized data such as genetic information and patient-reported health outcomes have even greater potential to address health disparities. However, exciting opportunities such as algorithms for machine learning using large amounts of data from smartphone health apps, or programs that analyze genetic mutations in a large patient population to unlock triggers for cancer growth, will require patients to consent to use of their data or ideally provide it themselves. Because of longstanding concerns about historical abuse at the hands of medical leaders as noted by Zhang et al, minority patients may be particularly suspicious about sharing their data with researchers or other entities. There are plenty of present-day concerns as well that justify these serious concerns – the number of individuals whose personal health information was affected by hacking/security breaches of network servers or EHRs jumped 10-fold from 2014 to 2015, exceeding 110 million people.[6] Health systems are already somewhat wary of sharing patient data to create the massive databases needed to advance this work due to a perceived loss of competitive advantage. At the same time, the potential for data breaches will further dampen enthusiasm if the scientific community cannot provide assurance to patients that their data will be safeguarded. Reversing each of these trends will require intensive efforts to demonstrate the value to patients of sharing their individual data and to health systems of sharing their aggregate data, as well as boosting data security safeguards using the most effective technology and eliminating all "weak links" within these shared data networks. These efforts are necessary in order for Big Data to usefully support clinical and population health research.

With access to the increased volume, velocity and variety offered by Big Data, we have to ask whether there is value in the use of these data to address health care disparities. In order to represent value, these data will need to support the investigation of fundamental questions such as: What are the source experiences, triggers or events that result in health disparities? What are the personal, health system and geographic barriers to improved individual health behaviors? What are the levers for action needed to address the identified health disparities? These data also can be used to deepen our understanding of heterogeneity and the implications for how we use Big Data. To the extent that "Models are opinions embedded in mathematics"[7] we will need to examine our embedded assumptions regarding the health care system we desire before we can conduct a complete and inclusive investigation of the health disparities that persist within this system. Big Data is not immune to these embedded assumptions. For example, if we assume that quality indices (and their contributing subdomains) can be applied to subpopulations in which they have not been formally tested, any errors in this assumption will be embedded in models that utilize Big Data to derive quality measures and then perpetuated when these measures are adopted across institutions.

The reduction and ultimate elimination of health disparities using Big Data is an admirable goal and one that we believe is achievable. Sustained progress toward this goal will require that we re-think how we generate, store, and protect patient data and will require the leadership to implement solutions to the existing challenges with Big Data. The articles represented in this issue advance the conversation by describing the current efforts, findings and concerns related to Big Data and health disparities. They offer important recommendations and perspectives to consider when designing systems that can usefully leverage Big Data to reduce health disparities. We hope that ongoing Big Data efforts can build on these contributions to advance the conversation, address our embedded assumptions, and identify levers for action to reduce health care disparities.

Acknowledgements

The contents of this article are those of the authors and do not necessarily represent the official view of the National Institutes of Health. E.P. is a staff member of the National Heart, Lung, and Blood Institute, National Institutes of Health.

References

1. United States Census. Tabulation and Processing. https://www.census.gov/history/www/innovations/technology/tabulation_and_processing.html Accessed February 14, 2017.
2. da Cruz, F. Columbia University Computing History. http://www.columbia.edu/cu/computinghistory/hollerith.html Accessed February 14, 2017.
3. Williams NB. The rise of Big Data: Trends and opportunity for the lab. Clinical Laboratory News. March 2014. https://www.aacc.org/publications/cln/articles/2014/march/big-data Accessed February 20, 2017.
4. Goudey, B, Abedini M, Hopper JL, et al. High performance computing enabling exhaustive analysis of higher order single

nucleotide polymorphism interaction in Genome Wide Association Studies. Health Inf Sci Syst, 2015. 3(Suppl 1 HISA Big Data in Biomedicine and Healthcare 2013 Con): S1-S3. https://doi.org/10.1186/2047-2501-3-S1-S3.

5. Yang C, Li C, Wang Q, Chung D, Zhao H. Implications of pleiotropy: challenges and opportunities for mining Big Data in biomedicine. *Frontiers in Genetics.* 2015 Jun 30;6:229. doi: 10.3389/fgene.2015.00229. eCollection 2015.
6. Office of the National Coordinator for Health Information Technology, Health IT Dashboard. https://dashboard.healthit.gov/quickstats/quickstats.php
7. Zhang, C. Q&A Cathy O'Neil, author of 'Weapons of Math Destruction,' on the dark side of big data. LA Times. December 30, 2016. http://www.latimes.com/books/jacketcopy/la-ca-jc-cathy-oneil-20161229-story.html Accessed: January 30, 2017.

Big Data articles in this issue of *Ethnicity & Disease*, Spring 2017

In the Rising Era of Big Data, Small Steps are Key
Nina Heller, Jonathan H. Seltzer

Racial and Ethnic Differences in a Linkage With the National Death Index
Eric A. Miller, Frances A. McCarty, Jennifer D. Parker

Hispanic Residential Isolation, ADHD Diagnosis and Stimulant Treatment among Medicaid-insured Youth
Dinci Pennap, Mehmet Burcu, Daniel J. Safer, Julie Magno Zito

Big Data Science: Opportunities and Challenges to Address Minority Health and Health Disparities in the 21st Century
Xinzhi Zhang, Eliseo J. Pérez-Stable, Philip Bourne, Emmanuel Peprah, Kenrik Duru, Nancy Breen, David Berrigan, Fred Wood, James S. Jackson, David W.S. Wong, Josh Denny

Enhancing Diversity in Biomedical Data Science
Judith E. Canner, Archana J. McEligot, María-Eglée Pérez, Lei Qian, Xinzhi Zhang