Big Data Science: Opportunities and Challenges to Address Minority Health and Health Disparities in the 21st Century

Xinzhi Zhang, MD, PhD[1]; Eliseo J Pérez-Stable, MD[1]; Philip E. Bourne, PhD[2]; Emmanuel Peprah, PhD[3]; O. Kenrik Duru, MD, MSHS[4]; Nancy Breen, PhD[1]; David Berrigan, PhD, MPH[5]; Fred Wood, DBA, MBA[6]; James S Jackson, PhD[7]; David W.S. Wong, PhD, MA[8]; Joshua Denny, MD, MS[9]

Addressing minority health and health disparities has been a missing piece of the puzzle in Big Data science. This article focuses on three priority opportunities that Big Data science may offer to the reduction of health and health care disparities. One opportunity is to incorporate standardized information on demographic and social determinants in electronic health records in order to target ways to improve quality of care for the most disadvantaged populations over time. A second opportunity is to enhance public health surveillance by linking geographical variables and social determinants of health for geographically defined populations to clinical data and health outcomes. Third and most importantly, Big Data science may lead to a better understanding of the etiology of health disparities and understanding of minority health in order to guide intervention development. However, the promise of Big Data needs to be considered in light of significant challenges that threaten to widen health disparities. Care must be taken to incorporate diverse populations to realize the potential benefits. Specific recommendations include investing in data collection on small sample populations, building a diverse workforce pipeline for data science, actively seeking to reduce digital divides, developing novel ways to assure digital data privacy for small populations, and promoting widespread data sharing to benefit under-resourced minority-serving institutions and minority researchers. With deliberate efforts, Big Data presents a dramatic opportunity for reducing health disparities but without active engagement, it risks further widening them.
*Ethn.Dis;*2017;27(2):95-106; doi:10.18865/ed.27.2.95.

Keywords: Big Data; Health Disparities; Health Inequities

Introduction

Although scientific and technological advances have improved the health and well-being of the US population overall, racial-ethnic minorities, socioeconomically disadvantaged, and other underprivileged or discriminated-against populations continue to experience a disproportionate share of many acute or chronic diseases and adverse health outcomes.[1-3] Big Data, defined by its volume, variety, velocity, variability, and veracity, is expected to bring significant benefits to health and health care, as it has to other sectors of the economy.[4,5] The improving quantity and quality of data, the changing dynamic and scale of data collection from various sources, and the fast development in measurements, analytic methods, and parallel computing of large amounts of biological and clinical data promise to dramatically transform clinical medicine and biomedical science. The growth of publicly traded companies in this arena suggests a belief in future profits in digital health care.[6,7] The question we address is, will the introduction of Big Data into clinical practice and health care research contribute to increasing health disparities or to decreasing them?

In March 2012, the US government announced the Big Data Research and Development Initiative.[8] Not long after, the National Institutes Health (NIH) established Big Data to Knowledge (BD2K), a trans-NIH initiative, to enable biomedical research to fully exploit the rich and massive digital research enterprise.[9]

[1]Office of the Director, National Institute on Minority Health and Health Disparities, National Institutes of Health (NIH)
[2]Office of the Director, NIH
[3]National Heart, Lung, and Blood Institute, NIH
[4]Department of Medicine, UCLA Medical Center
[5]National Cancer Institute, NIH
[6]National Library of Medicine, NIH
[7]College of Literature, Science and the Arts, University of Michigan
[8]Department of Geography and GeoInformation Science, College of Science, George Mason University
[9]Departments of Biomedical Informatics and Medicine, Vanderbilt University Medical Center

Address correspondence to Xinzhi Zhang, MD, PhD; Program Director; Division of Extramural Scientific Programs (DESP); NIMHD, NIH; 6707 Democracy Boulevard, Suite 800; Bethesda, Maryland 20892-5465. 301.594.6136; xinzhi.zhang@nih.gov

In 2016, the National Institute on Minority Health and Health Disparities (NIMHD) held a workshop on methods and measurements science and concluded that addressing minority health and health disparities research had been missing from these Big Data science initiatives, and that leadership was needed in rectifying this deficiency. With NIMHD's leadership, NIH Institutes and Centers and other federal agencies are starting to utilize Big Data to address health disparities. In recognition of the need to address disparities, the NIH Precision Medicine Initiative's one-million-person cohort has an explicit focus on diversity, and will target recruitment of historically understudied populations.[10] These events suggest that the time is right to leverage the growing impetus in Big Data for the purposes of reducing health disparities. Here, we outline how Big Data may be used to advance understanding of minority health and reduce health disparities and list recommendations for moving the field forward.

Opportunities to Use Big Data Science to Advance Minority Health and Reduce Health Disparities

From data to information, from information to knowledge, and finally from knowledge to evidence-based practice, Big Data is changing medical practice and public health.[11] Understanding health disparities requires understanding the interactions of influences that shape health disparities at various levels (individual, interpersonal, family, community, societal) over the life course, the diversity of the relevant mediators (exposures, resiliency), and the multiple interacting mechanisms involved (biological, socioeconomic, behavioral, and environmental). The ecosystem of Big Data comprises multimodal, multifactorial, and multilevel data sources for data mining, and potentially provides the environment to both study and address health disparities. The challenge is to ensure that the promise of Big Data will be realized to increase access to health care and improve health promotion and quality of health care for disadvantaged and discriminated-against groups so that minority health improves and disparities are reduced. There are three priority opportunities that Big Data science may offer to the reduction of health and health care disparities. One opportunity is to incorporate standardized information on demographic and social determinants in electronic health records in order to target ways to improve quality of care for the most disadvantaged populations over time. A second opportunity is to enhance public health surveillance by linking geographical variables and social determinants of health for geographically defined populations to clinical data and health outcomes. Third and most importantly, Big Data science may lead to a better understanding of the etiology of health disparities and understanding of minority health to guide intervention development.

We outline how Big Data may be used to advance understanding of minority health and reduce health disparities and list recommendations for moving the field forward.

Opportunity I: To Incorporate Social Determinants Information and Improve Quality of Care for Underserved Populations

The HITECH Act[12] spurred adoption of electronic health records (EHRs) throughout the United States. The vast majority of health care systems now have EHRs.[13,14] This growth in EHRs is the foundation for Big Data in health and medicine and could be a foundation for reducing health disparities. Importantly, this increase in adoption in EHRs has been seen across both large nonprofit and for-profit health care institutions, individual and small clinical practices and Federally Qualified Health Centers (FQHC),[15] creating new opportunities to study health disparities populations whose medical data were not previously available in electronic format. It also provides the first opportunity to incorporate information on standardized demographic and social determinants of health on a large scale. The resulting data would allow social needs to be addressed in clinical settings and the underlying causes of health disparities to be understood.

The Institute of Medicine's Committee on Recommended Social and Behavioral Domains and Measures for Electronic Health Records has identified selected domains and measures that capture the social determinants of health to inform the development of recommendations for meaningful use of EHRs.[16,17] In 2015, 96% of non-federal acute care hospitals possessed a certified EHR technology adopted by the Department of Health and Human Services and 84% have at least a basic EHR system.[18] At FQHC sites, 80% use EHRs and 75% have demonstrated meaningful use of certified EHR technology.[15] However, differences exist between EHR systems in large, well-resourced clinical practices compared with less well-resourced FQHC sites in their ability to support population health management and "meaningful use" to track and address disparities.[18,19] If patient, family, and community focus and shared decision making were implemented equally in these two types of settings, social determinants of health information could both improve public health,[16,17,20] and reduce the disparities that otherwise would arise with the adoption of Big Data technology.

Big Data relevant to health and health care can encompass clinical registries, lab tests, diagnoses, and medications in the EHRs, insurance claims data, medical imaging, biobanks, genomic sequence data, Food and Drug Administration (FDA)'s safety monitoring data, biometric data from consumer grade appliances, or population surveys. Big Data and information technology hold out potential for the health care industry to improve quality of care, reduce unnecessary cost, and promote prevention and healthy lifestyles for the population; however, vigilance will be needed to ensure that it does not also generate greater disparities by contributing to the digital divide. The impact of technology has left behind minority and low socioeconomic status (SES) populations in the past and we need to guard against this with the inception of Big Data collection, analytics, and associated technologies. For example, selected health indicators can be utilized to assess whether minority and health disparity populations receive the same quality of care as other populations. Large clinical registries based on EHRs may be used to assess different treatment strategies, analyze longitudinal outcomes and adverse effects for large cohorts of diverse patients, and capture uncommon diseases or conditions that are rarely examined in traditional clinical trials. However, analyzing these data is not easy due to differences in EHR encoding systems, and data fragmentation across practices and institutions. Networks such as the Electronic Medical Records and Genomics (eMERGE) network have been addressing these challenges.[21]

Big Data provides an opportunity for personalized care for everyone and may be used in precision medicine to optimize treatment for individual patients.[22,23] It has the potential to especially benefit racial-ethnic minority and other underserved populations for whom we do not have evidence, because most clinical trial data were analyzed without adequate numbers of minority or low SES populations.[24] With the adoption of EHRs in all health care settings, and the incorporation of additional digital health information from monitoring, big clinical data will be generated and available to provide the means for conducting pragmatic trials including underserved populations and to help compensate for the lack of disparity populations in randomized clinical trials.[25] In combination with large-scale cost data, clinical outcome data can also be useful to conduct comparative effectiveness and cost-effectiveness analysis to inform medical decision making and policy on appropriate coverage of tests and medications.[26] Nevertheless, this potential will only be realized with accrual of Big Data across diverse populations using standardized categories. A challenge will be to include all Americans in health care delivery so records are available to improve their quality of care. Importantly, there needs to be a concerted effort to apply precision medicine to address issues of minority health and health disparities right from the beginning.

Opportunity II: To Improve Public Health Surveillance and Address Health Disparities

The expanded access to health care under the implementation of the Affordable Care Act (ACA) has significantly benefitted racial-ethnic minorities and people with low SES.[27] For instance, the percentage of uninsured Latino adults aged 18-64 have decreased from 40.6% in 2013 to 28.3% in 2015.[28] There was a significant decrease in the percentage of uninsured adults after the ACA, most dramatically among adults who were poor (<100% federal poverty level

[FPL]; from 39.3% in 2013 to 28.0% in 2015) or near-poor (≥100% and <200% FPL; from 38.5% in 2013 to 23.8% in 2015).[28] ACA improved coverage for preventive and treatment services. This benefits the millions of underserved Americans who could not afford preventive services if copayment was required.[29] More expansive insurance coverage for a larger percentage of the population, especially persons with chronic diseases, may generate additional EHR data that is more representative, including populations that are more likely to experience disparities.

Generation of clinical and other Big Data resources related to health over time and combining it with environmental and policy data collected prospectively, could allow spatiotemporal surveillance and monitoring systems in different micro-environments (eg, combinations of EHRs, local public health clinics, communities, and political units). Evaluation of these data would identify areas with disparities, whether disparities are decreasing or increasing, and the factors associated with disparities. Factors closely associated with disparities could be used to identify areas at risk for disparities. The availability of large amounts of health disparities data in a national surveillance system would make it possible for monitoring and tracking burden and trends of disparities. The FDA Sentinel System is a national electronic surveillance system for medical devices to track adverse events and assess safety.[30] Combining EHR data with FDA reporting systems, molecular data, and/or social media has identified potential drug-drug interactions and side effects.[31,32] In addition, millions of clinical notes from EHRs could be mined to systematically monitor post-marketing adverse drug events.[33] Efforts should be made to use these systems to address disparities reduction.

Big Data can be used to assess national and local public health policies and other natural experiments to promote health and prevent diseases. For example, the National Health Interview Survey is used to estimate insurance coverage in different segments of the US population, and clinical data are being used to measures access- and quality-related outcomes. Visualization and network analysis techniques that have emerged with Big Data offer opportunities to link community-level data with health care system data. Use of these techniques on Big Data would enable public health officials and clinicians to more efficiently allocate resources and to assess whether all patients are getting the medical services they need. Geographic information systems (GIS) can be used to locate social determinants of health and help focus public health interventions on populations at greater risk of health disparities. For example, Duke University used GIS to visualize the distribution of individuals with diabetes across Durham County, NC. GIS was used to explore gaps in access to care and self-management resources and to direct resources into areas of need.[34] Place-based health disparities is emerging as an important area of research that can inform future policy.[3,35] Social media data hold the promise of linking social context to health/well-being and behavioral change. Such linkages could help identify the social contexts that lead to reduction of disparities.

Novel technologies may be able to identify place-based disparities in chronic diseases and epidemics. For example, Young, Rivers, and Lewis analyzed 553,186,061 tweets and found a significant association between the geographic locations of HIV-related tweets and HIV prevalence,[36] which provided epidemiological evidence for future targeted community-level interventions and surveillance using Twitter. Google used Big Data generated by search requests to identify or forecast the location of flu epidemics by analyzing associated Internet Protocol (IP) addresses,[37] although the results were later withdrawn after extensive public reaction.[38] Given that minority populations are historically less likely to access preventive services, such geographic information identified from social media data may especially benefit minority and low SES populations during future emergency responses including stockpiling for pandemic influenza.

Opportunity III: To Understand Etiology and to Guide Interventions to Reduce Disparities

Not all clinical research questions can be studied or tested in randomized controlled trials due to scientific, operational, ethical, or cost concerns; using Big Data in simulation modeling and systems science provides an opportunity to model data in response to challenging questions that offer insight on how to address them. Simulation modeling is especially useful for minority health and health disparities research because it can model system-

ic and ecological causes that accumulate over the life course. Modeling can also test whether interventions are scalable and sustainable using a multidisciplinary, community-engaged approach.[39] In a systematic review of simulation models for socioeconomic inequalities, Speybroeck et al concluded that agent-based modeling, a powerful simulation modeling technique, is an appropriate tool for examining health disparities because it can simulate the complex nature of health inequalities.[40] Big Data simulation modeling has the potential to be more accurate than traditional modeling techniques, especially when ample individual and institution-level information connected and harmonized from various sources are available.[41] Big Data simulation modeling could potentially accelerate the progress in determining the relative importance of different causal factors of health disparities, which may not be feasible in observational studies.

Predictive modeling has used clinical data in various situations to forecast probable complications and guide clinical decision making.[42,43] Early detection of high-risk patients can lead to early diagnosis and early intervention that may lead to better health outcomes and cost savings. In many cases, the burden from the more severe stages of the disease disproportionately affects minority patients and those with low SES, and therefore, early diagnosis and timely treatment may provide greater benefit for those populations subject to worse outcomes. For example, machine learning applied to clinical data has been used to predict acute care use and cost of treatment for asthmatic patients, diagnose diabetes among adults, predict in-hospital mortality and drug response, improve disease classification, and identify disease subsets.[44-47] Taylor et al suggest that a machine learning algorithm using Big Data conforms to actual real time clinical practice, allows incorporation of far more clinical variables, and may assist in discovering unexpected predictors.[48] Big Data analytic tools such as natural language processing, machine-learning, or electronic case-finding algorithms applied to EHR data have produced a number of insights into genomics of disease and drug response.[32] Some of these findings may explain apparent disparities in care, such as poorer response to clopidogrel in Pacific Islanders[49] or higher doses of tacrolimus required for African Americans (which can lead to under-dosing and thus increased risk of acute transplant rejection).[50] Future use of such methods applied to massive datasets of EHRs and other data may help identify disparities populations at high risk of chronic diseases (eg, cardiovascular diseases, diabetes, and asthma) or infectious diseases (eg, influenza, hepatitis) and address risk factors through timely interventions (eg, obesity/diabetes prevention, vaccination).[43]

Potential Challenges of Using Big Data for Minority Health and Health Disparities Research

Although many potential challenges of Big Data are applicable to all research studies, these challenges may have a more adverse impact on minority health and health disparities research. Although 74.4% of households reported having broadband access to the Internet in 2013, disparities in access to Internet and health information remain.[51] Data from 2011 National Health Interview Survey reported that Whites were more likely to use the Internet to search for health information compared with other races/ethnicities and the percentage of adults who search for health information increased with education level.[52]

The promise of Big Data may be offset by challenges that threaten to widen health disparities. Moreover, persons with a more disadvantaged status are particularly vulnerable to unintended adverse effects of information system transformations. Specific recommendations include investing in data collection on small sample populations, building a diverse workforce pipeline for data science, actively seeking to reduce digital divides, developing novel ways to assure digital data privacy for small populations, and promoting widespread data sharing to benefit under-resourced minority-serving institutions and minority researchers.

Challenge I: Ethics, Privacy, and Trust

A key advantage of Big Data analytics is through linking disparate data sources, which requires access to personal identifiable information (PII) or at least some proxy.[53] Use of PII presents privacy and ethical concerns.[54] One way to protect privacy while sharing PII is to use privacy-preserving data linkage

models, which share collections of one-way hashed identifiers to align diverse datasets.[55] However, these systems require both datasets to have access to PII (or pre-hashed identifiers), and many current potential data providers may not have the ability at this time to implement such a system due to technical and cost reasons. Data de-identification can help mitigate privacy concerns. However, even data that is de-identified according to standards such as Safe Harbor are not necessarily anonymous – since unique de-identified data can be re-identifiable by triangulation across other data sources.[56,57] Public data from Google or Twitter can point to an individual IP address, location, or other personal information and may require additional layers of oversight. Informed consent or assent for traditional clinical trials or studies may not be applicable for analyses of Big Data with potential personal information that imposes new challenges for Institutional Review Boards (IRB). Given the complicated situation, the White House report on Big Data and privacy called for regulations that focus on the use of data via providers rather than trying to regulate collection or analysis of data.[58] Privacy concerns will need to be addressed for widespread data linkage to occur.

Developing Trust in the System

Loss of confidentiality or misuse of sensitive personal information can endanger the individual patient. A particular issue in health disparities research is lack of trust that has evolved in health care because of unethical treatment of disenfranchised minority populations. The Tuskegee Study of Untreated Syphilis,[59] the Henrietta Lacks case,[60] and the diabetes studies of the Pima Indians[61] are examples that have created mistrust in US health care and scientific institutions. Mistrust of the health care system by entire population groups has led to an increased emphasis by researchers on community engagement and participation in health disparities research. This same credo is crucial to ensure that Big Data science serves minority populations in a respectful and beneficial way. Minority-serving institutions usually do not have the infrastructure that research-intensive universities have to capture, manage and analyze Big Data. Collaborations between minority-serving institutions and research-intensive institutions are needed to take advantage of the rapid growth of health informatics and technologies such that they will lead to the reduction of health disparities.

Avoiding "Cherry-Picking" Patients

EHR systems focusing on quality metrics may be used to identify high health care utilizers and patients with serious medical conditions or living in social disadvantage, so that health care systems and clinicians may provide better care with available resources.[42] However, these Big Data analytics may lead to a greater digital divide and be used to avoid the high costs associated with serving patient populations who are more likely to be from minority groups or poor.[62] As a consequence, these patients may be encouraged to seek less appropriate care, be declined needed services or referrals, sent to a safety-net clinical system for care, or be asked for higher out-of-pocket payment. Terms like "frequent flyers" used in emergency departments and psychiatric crisis centers to identify high health care utilizers demonstrate implicit bias.[63] Clinicians and other staff should be sensitive to ethical iconography and language. How to ensure equal access and equal quality of service remains a topic to be addressed. To avoid cherry-picking patients and rebuild trust among minority or disadvantaged populations, legislative protection and regulation assurance are warranted. From a population health perspective, using Big Data to evaluate quality of care and ensure excellent care of the most vulnerable patients in a health care system should be one of the metrics of value-based care.

Challenge II: Missing Data and Statistical Uncertainty

Well-analyzed Big Data can bring novel insights but poorly analyzed Big Data can be misinterpreted, especially in minority health and health disparities research, where results lacking social or cultural context can be misleading.[64] Existing EHRs may not have good quality data on health disparities related information, including missing socioeconomic information and institutional variability on data standards. Progress in health disparities research and science will require improvements in the completeness, standardization and validity of demographic measures and social de-

terminants reported from multiple sources, including electronic medical records, clinical trials, genomic research, and various forms of administrative records such as Medicare and Medicaid. Other types of data sources such as surveys, extrapolations, and imputations may suffice for national reports and overall trending, but are insufficient for analyzing places which, as we have seen, is critical for health disparities research. Further, health disparities populations must be fully incorporated in the precision medicine cohort and research questions and in similar cutting edge personalized biomedical initiatives.[65,66]

Statistical uncertainty may still be a problem when data are "big."[67] Small differences in Big Data may be statistically significant because of the large number of observations, but the findings may not be useful for clinicians or patients.[68] Moreover, conclusions drawn from Big Data cannot automatically be generalized to minority populations. Uncertainty around these issues related to Big Data may be resolved in the future with newly developed methods, algorithms, technologies, and sound statistical training; however, this will not happen unless health disparities research is a consistent focus in the development of Big Data. Another concern is that Big Data may not collect race/ethnicity or may overlook certain small sample populations (eg, American Indians, Alaska Natives, Pacific Islanders, and sexual and gender minorities) with unique characteristics that may be critical for understanding etiology of specific conditions and health care delivery in such populations.[69]

Challenge III: Data Access and Sharing

The power of Big Data cannot be achieved unless challenges such as secure storage, integration, harmonization, access, and sharing are addressed.[70] Data sharing is essential for translating research findings to improve human health. The NIH policy requires that research data be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data.[71] However, much of available data may be proprietary or protected, eg, such as falling under Health Insurance Portability and Accountability Act (HIPAA), and would require novel approaches and/or individual consent to access. Interactive data retrieval is a critical component for data sharing and data security.[72] To address the lack of data interoperability standards, Bahga and Madisetti propose a cloud-based approach for the design of interoperable EHR systems for clinicians, patients, and third-party payers.[73] Systems like MedCloud[74] and Home-Diagnosis[75] were proposed to manage large patient data and for conducting analysis. However, articulation of EHR data can be challenging as different EHR systems may use custom-made ("bespoke") encoding systems and variable names.[76] To handle this, common data models such as Observational Medical Outcomes Partnership (OMOP),[77] PCORnet,[78] and the Shared Health Research Informatics Network (SHRINE) implementation of i2b2 (Informatics for Integrating Biology and the Bedside)[79] have been proposed. Acceptance of strategies to address these problems is gaining ground, but conversions to common data models are not trivial.

Doshi et al reviewed the access policies of publicly funded patient-level clinical data and concluded that removal of unnecessary barriers to utilization of these valuable resources were needed.[80] They suggested placing more emphasis on research quality and less on which institution the researcher belongs to; encouraging more identifiable research information and data linkage; promoting easy remote access; and implementing tiered pricing for data usage fees. These recommendations may reduce data-access disparities among researchers. Additionally, the medical research community was urged to consider novel approaches to share data including non-positive findings.[81] Although this issue is not specific to health disparities research, minority scientists especially those in under-resourced institutions, are more likely to experience such barriers and may benefit more from open data policies.

Challenge IV: Data Science Training and Workforce Diversity

Big Data science brings together clinicians, health researchers, government agencies, commercial enterprises, and patients in one place for information exchange. Data scientists will need to partner with physicians, nurses, researchers, as well as patients to better understand the data and transform unstructured or structured numbers into systemic information and knowledge. In the future, patient consumers of Big Data

Table 1. Minority health and health disparities relevant recommendations on Big Data science

1. Incorporate standardized collection and input of race/ethnicity, socioeconomic status and other social determinants of health measures in all systems that collect health data (Opportunity I)
2. Enhance public health surveillance by incorporating geographical variables and social determinants of health for geographically defined populations (Opportunity II)
3. Advance simulation modeling and systems science using big data to understand the etiology of health disparities and guide intervention development (Opportunity III)
4. Build trust to avoid historical concerns and current fears of privacy loss and "big brother surveillance" through sustainable long-term community relationships (Challenge I)
5. Invest in data collection on area relevant small sample populations to address incompleteness of big data (Challenge II)
6. Encourage data sharing to benefit under-resourced minority-serving institutions and underrepresented minority researchers by research intensive institutions (Challenge III)
7. Promote data science in training programs for underrepresented minority scientists (Challenge IV)
8. Assure active efforts are made up front during both the planning and implementing stages of new big data resources to address disparities reduction (Challenges I-IV)

may demand specific clinical trials, individualized treatment plans, and precision or personalized medication.

According to the biennial report "Women, Minorities, and Persons with Disabilities in Science and Engineering," mandated by the Science and Engineering Equal Opportunities Act (Public Law 96-516), the gap in educational attainment separating underrepresented minorities from Whites and Asians remains wide in mathematics, statistics, and computer sciences.[82] Both underrepresentation of investigators from diverse racial and ethnic minority populations and persistent health disparities warrant the urgent need for policies to improve scientific workforce diversity in the United States.[83] Lack of resources to process large amounts of data and to perform more sophisticated data mining and statistical analysis have limited the education and training opportunities of underrepresented students. This is especially true for students who are educated and trained in resource-limited universities, which lack access to an informatics infrastructure with high power computing capabilities. This leads to disadvantages in seeking funding and other support. Thus, training and education of underrepresented students and faculty, as well as providing resources to minority-serving and other under-resourced universities, is a critical component of the Big Data enterprise.[84]

To realize the potential of Big Data, a focus on health disparities is needed during the planning and implementing of Big Data resources.

NIH has acknowledged the needs for data science training and established the BD2K Diversity program[85] in some resource-limited institutions such as California State University (Northridge and Monterey Bay), Fisk University, and University of Puerto Rico. Such efforts will bring advanced data science technology and skill sets to underrepresented minority students and eventually build a diverse data science pipeline for future generations.[86]

Conclusion

In the era of information explosion, Big Data approaches are likely to be able to contribute to understanding the causes of health disparities and to identifying useful opportunities for their reduction, but only if Big Data collection includes health disparities populations and if researchers who focus on these populations are trained to use Big Data. Big Data could lead to new discoveries and new experiments in health disparities research that were never before possible. To realize this potential, a focus on health disparities is needed during the planning

and implementing of Big Data resources. Otherwise, it is likely that these promising new approaches will worsen disparities. Table 1 presents a list of recommendations highlighting the opportunities and challenges of Big Data science to address minority health and health disparities in the 21st century.

As Big Data is collected, all facets of the US population need to be represented to accurately describe the health of the US population and to understand the etiology of health disparities. This scientific foundation is needed to address disparities. Big Data can enhance public health surveillance by incorporating geographical variables and social determinants of health. Big Data promises accurate and standardized measurement of exposures, outcomes, and confounders, which are critical to analyzing health disparities. Simulation modeling with Big Data holds promise for understanding the causes of health disparities and guiding the development and implementation of interventions. Finally, investments are needed to: build trust; avoid historical mistakes; protect privacy; ensure systematic data collection that represents all segments of the populations including small sample populations; make available data sharing that will benefit under-resourced minority-serving institutions and minority researchers; and develop a diverse workforce pipeline for data science. With deliberate efforts, Big Data presents an effective opportunity to reduce health disparities; however, without active engagement, disparities are likely to widen.

Acknowledgements

This article was funded in part by the National Institutes of Health, Office of the Director, National Institute on Minority Health and Health Disparities, National Cancer Institute, and National Library of Medicine intramural and/or extramural programs. We also thank the NIH BD2K executive committee and BD2K program management working groups for unstinting support.

Conflict of Interest

No conflicts of interest to report.

Author Contributions

Research concept and design: Zhang, Pérez-Stable, Bourne, Berrigan, Wood, Jackson, Wong, Denny; Acquisition of data: Zhang, Pérez-Stable, Berrigan; Data analysis and interpretation: Zhang, Pérez-Stable, Peprah, Duru, Breen, Berrigan; Manuscript draft: Zhang, Pérez-Stable, Bourne, Peprah, Duru, Breen, Berrigan, Wood, Jackson, Wong, Denny; Statistical expertise: Zhang, Breen, Wong; Acquisition of funding: Pérez-Stable; Administrative: Zhang, Bourne, Peprah, Duru, Berrigan, Wood, Jackson, Wong; Supervision: Pérez-Stable, Denny

References

1. Ayanian JZ, Landon BE, Newhouse JP, Zaslavsky AM. Racial and ethnic disparities among enrollees in Medicare Advantage plans. *N Engl J Med*. 2014;371(24):2288-2297. https://doi.org/10.1056/NEJMsa1407273. PMID:25494268.
2. Manrai AK, Funke BH, Rehm HL, et al. Genetic Misdiagnoses and the Potential for Health Disparities. *N Engl J Med*. 2016;375(7):655-665. https://doi.org/10.1056/NEJMsa1507092. PMID:27532831.
3. Dankwa-Mullan I, Pérez-Stable EJ. Addressing Health Disparities Is a Place-Based Issue. *Am J Public Health*. 2016;106(4):637-639. https://doi.org/10.2105/AJPH.2016.303077. PMID:26959267.
4. Berman JJ. *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*. Amsterdam, Netherlands : Elsevier, 2013.
5. Hilbert M. Big Data for development: a review of promises and challenges. *Dev Policy Rev*. 2016;34(1):135-174. https://doi.org/10.1111/dpr.12142.
6. Hoyt RE, Snider D, Thompson C, Mantravadi S. IBM Watson analytics: automating visualization, descriptive, and predictive statistics. *JMIR Public Health and Surveillance.* 2016;2(2):e157.
7. Chen Y, Elenee Argentinis JD, Weber G. IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clin Ther*. 2016;38(4):688-701. https://doi.org/10.1016/j.clinthera.2015.12.001. PMID:27130797.
8. The White House. *Big Data is a Big Deal.* U.S. Government; 2012. https://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal. Accessed October 17, 2016.
9. National Institutes of Health. *Data Science at NIH.* National Institutes of Health; 2012. https://datascience.nih.gov/. Accessed October 17, 2016.
10. Precision Medicine Initiative (PMI) Working Group. *The Precision Medicine Initiative Cohort Program – Building a Research Foundation for 21st Century Medicine.* National Institutes of Health;2015. https://www.nih.gov/sites/default/files/research-training/initiatives/pmi/pmi-working-group-report-20150917-2.pdf: Accessed October 17, 2016.
11. Sim I. Two ways of knowing: Big Data and evidence-based medicine. *Ann Intern Med*. 2016;164(8):562-563. https://doi.org/10.7326/M15-2970. PMID:26809201.
12. Pipersburgh J. The push to increase the use of EHR technology by hospitals and physicians in the United States through the HITECH Act and the Medicare incentive program. *J Health Care Finance*. 2011;38(2):54-78. PMID:22372032.
13. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving electronic medical records upstream: incorporating social determinants of health. *Am J Prev Med*. 2015;48(2):215-218. https://doi.org/10.1016/j.amepre.2014.07.009. PMID:25217095.
14. Adler NE, Stead WW. Patients in context--EHR capture of social and behavioral determinants of health. *N Engl J Med*. 2015;372(8):698-701. https://doi.org/10.1056/NEJMp1413945. PMID:25693009.
15. Office of the National Coordinator for Health Information Technology. *Percent of REC Enrolled Providers in an Organization/Site and Area Type Live on an EHR and Demonstrating Meaningful Use.* Washington, DC; 2016.
16. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records; Board on Population Health and Public Health Practice. *Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1*. Washington, DC: Institute of Medicine. 2014.
17. Committee on the Recommended Social and Behavioral Domains and Measures

for Electronic Health Records; Board on Population Health and Public Health Practice. *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. Washington, DC: Institute of Medicine. 2015.
18. Henry J, Pylypchuk Y, Searcy T, Patel V. *Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015.* Washington, DC: Office of the National Coordinator for Health Information Technology; 2016.
19. Kruse CS, Kothman K, Anerobi K, Abanaka L. Adoption Factors of the Electronic Health Record: A Systematic Review. *JMIR Med Inform*. 2016;4(2):e19. https://doi.org/10.2196/medinform.5525. PMID:27251559.
20. Garg A, Boynton-Jarrett R, Dworkin PH. Avoiding the unintended consequences of screening for social determinants of health. *JAMA*. 2016;316(8):813-814. https://doi.org/10.1001/jama.2016.9282. PMID:27367226.
21. Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genetics in Med.* 2013;15(10):761-771.
22. Prentice JC, Conlin PR, Gellad WF, Edelman D, Lee TA, Pizer SD. Capitalizing on prescribing pattern variation to compare medications for type 2 diabetes. *Value in Health.* 2014;17(8):854-862.
23. Hripcsak G, Ryan PB, Duke JD, et al. Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci USA*. 2016;113(27):7329-7336. https://doi.org/10.1073/pnas.1510502113. PMID:27274072.
24. Chen MS Jr, Lara PN, Dang JHT, Paterniti DA, Kelly K. Twenty years post-NIH Revitalization Act: enhancing minority participation in clinical trials (EMPaCT): laying the groundwork for improving minority clinical trial accrual: renewing the case for enhancing minority participation in cancer clinical trials. *Cancer*. 2014;120(suppl 7):1091-1096. https://doi.org/10.1002/cncr.28575. PMID:24643646.
25. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol*. 2016;183(8):758-764. https://doi.org/10.1093/aje/kwv254. PMID:26994063.
26. Collins B. Big Data and Health Economics: Strengths, Weaknesses, Opportunities and Threats. *Pharmacoeconomics*. 2016;34(2):101-106. https://doi.org/10.1007/s40273-015-0306-7. PMID:26093888.
27. Obama B. United States Health Care Reform: Progress to Date and Next Steps. *JAMA*. 2016;316(5):525-532. https://doi.org/10.1001/jama.2016.9797. PMID:27400401.
28. National Center for Health Statistics, Centers for Disease Control and Prevention. Health Insurance Coverage: Early Release of Estimates from the National Health Interview Survey, January–March 2015. http://www.cdc.gov/nchs/data/nhis/earlyrelease/insur201508.pdf. 2016. Accessed Feb 15, 2017.
29. Bergner L, Yerby AS. Low income and barriers to use of health services. *N Engl J Med*. 1968;278(10):541-546. https://doi.org/10.1056/NEJM196803072781006. PMID:4866348.
30. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative--A comprehensive approach to medical product surveillance. *Clin Pharmacol Ther*. 2016;99(3):265-268. https://doi.org/10.1002/cpt.320. PMID:26667601.
31. Tatonetti NP, Denny JC, Murphy SN, et al. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. *Clin Pharmacol Ther*. 2011;90(1):133-142. https://doi.org/10.1038/clpt.2011.83. PMID:21613990.
32. White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. *J Am Med Inform Assoc*. 2013;20(3):404-408. https://doi.org/10.1136/amiajnl-2012-001482. PMID:23467469.
33. Wang G, Jung K, Winnenburg R, Shah NH. A method for systematic discovery of adverse drug events from clinical notes. *J Am Med Inform Assoc*. 2015;22(6):1196-1204. https://doi.org/10.1093/jamia/ocv102. PMID:26232442.
34. Spratt SE, Batch BC, Davis LP, et al. Methods and initial findings from the Durham Diabetes Coalition: integrating geospatial health technology and community interventions to reduce death and disability. *J Clin Transl Endocrinol*. 2015;2(1):26-36. https://doi.org/10.1016/j.jcte.2014.10.006.
35. Linton SL, Cooper HL, Kelley ME, et al; National HIV Behavioral Surveillance Study Group. Associations of place characteristics with HIV and HCV risk behaviors among racial/ethnic groups of people who inject drugs in the United States. *Ann Epidemiol*. 2016;26(9):619-630.e2. https://doi.org/10.1016/j.annepidem.2016.07.012. PMID:27576908.
36. Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Prev Med*. 2014;63:112-115. https://doi.org/10.1016/j.ypmed.2014.01.024. PMID:24513169.
37. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457(7232):1012-1014. https://doi.org/10.1038/nature07634. PMID:19020500.
38. Lazer D, Kennedy R, King G, Vespignani A. Big Data. The parable of Google Flu: traps in Big Data analysis. *Science.* 2014;343(6176):1203-1205.
39. Smith BT, Smith PM, Harper S, Manuel DG, Mustard CA. Reducing social inequalities in health: the role of simulation modelling in chronic disease epidemiology to evaluate the impact of population health interventions. *J Epidemiol Community Health*. 2014;68(4):384-389. https://doi.org/10.1136/jech-2013-202756. PMID:24363409.
40. Speybroeck N, Van Malderen C, Harper S, Müller B, Devleesschauwer B. Simulation models for socioeconomic inequalities in health: a systematic review. *Int J Environ Res Public Health*. 2013;10(11):5750-5780. https://doi.org/10.3390/ijerph10115750. PMID:24192788.
41. Gange SJ, Golub ET. From Smallpox to Big Data: The Next 100 Years of Epidemiologic Methods. *Am J Epidemiol*. 2016;183(5):423-426. https://doi.org/10.1093/aje/kwv150. PMID:26443419.
42. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff (Millwood)*. 2014;33(7):1123-1131. https://doi.org/10.1377/hlthaff.2014.0041. PMID:25006137.
43. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol*. 2016;13(6):350-359. https://doi.org/10.1038/nrcardio.2016.42. PMID:27009423.
44. Luo G. PredicT-ML: a tool for automating machine learning model building with big clinical data. *Health Information Science and Systems.* 2016;4:5.
45. Carroll RJ, Eyler AE, Denny JC. Naive Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA Symposium Proceedings*. 2011;2011:189-196.
46. Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*. 2014;133(1):e54-e63. https://doi.org/10.1542/peds.2013-0819. PMID:24323995.
47. Peissig PL, Santos Costa V, Caldwell MD, et al. Relational machine learning for electronic health record-driven phenotyping. *J Biomed Inform*. 2014;52:260-270. https://doi.org/10.1016/j.jbi.2014.07.007.

PMID:25048351.
48. Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Academic Emerg Med.* 2016;23(3):269-278.
49. Wu AHB, White MJ, Oh S, Burchard E. The Hawaii clopidogrel lawsuit: the possible effect on clinical laboratory testing. *Per Med*. 2015;12(3):179-181. https://doi.org/10.2217/pme.15.4.
50. Beermann KJ, Ellis MJ, Sudan DL, Harris MT. Tacrolimus dose requirements in African-American and Caucasian kidney transplant recipients on mycophenolate and prednisone. *Clin Transplant*. 2014;28(7):762-767. https://doi.org/10.1111/ctr.12376. PMID:24754564.
51. File T, Ryan C. *Computer and Internet Use in the United States: 2013.* Washington, DC: American Community Survey Reports, ACS-28, U.S. Census Bureau; 2014.
52. Amante DJ, Hogan TP, Pagoto SL, English TM, Lapane KL. Access to care and use of the Internet to search for health information: results from the US National Health Interview Survey. *J Med Internet Res*. 2015;17(4):e106. https://doi.org/10.2196/jmir.4126. PMID:25925943.
53. Mittelstadt BD, Floridi L. The ethics of big data: current and foreseeable issues in biomedical contexts. *Sci Eng Ethics*. 2016;22(2):303-341. https://doi.org/10.1007/s11948-015-9652-2. PMID:26002496.
54. Amir Y, Sharon I. Replication research - a must for the scientific advancement of psychology. *J Soc Behav Pers*. 1990;5(4):51-69.
55. Kho AN, Cashy JP, Jackson KL, et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. *J Am Med Inform Assoc*. 2015;22(5):1072-1080. https://doi.org/10.1093/jamia/ocv038. PMID:26104741.
56. Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. *J Am Med Inform Assoc*. 2010;17(3):322-327. https://doi.org/10.1136/jamia.2009.002725. PMID:20442151.
57. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science*. 2013;339(6117):321-324. https://doi.org/10.1126/science.1229566. PMID:23329047.
58. The President's Council of Advisors on Science and Technology. *Report to the President: Big Data and Privacy: A Technological Perspective.* Washington, DC: White House; 2014.
59. White RM. Unraveling the Tuskegee Study of Untreated Syphilis. *Arch Intern Med*. 2000;160(5):585-598. https://doi.org/10.1001/archinte.160.5.585. PMID:10724044.
60. Caplan A. *NIH finally makes good with Henrietta Lacks' family -- and it's about time, ethicist says.* NBC News; http://www.nbcnews.com/health/nih-finally-makes-good-henrietta-lacks-family-its-about-time-6C10867941; 2013. Accessed October 17, 2016.
61. Young E. *Making Indigenous Peoples Equal Partners in Gene Research.* http://www.theatlantic.com/science/archive/2015/10/indigenising-genomics/412096/2015. Accessed October 17, 2016.
62. Wears RL, Williams DJ. Big questions for "Big Data". *Ann Emerg Med*. 2016;67(2):237-239. https://doi.org/10.1016/j.annemergmed.2015.09.019. PMID:26481264.
63. Joy M, Clement T, Sisti D. The ethics of behavioral health information technology: frequent flyer icons and implicit bias. *JAMA*. 2016;316(15):1539-1540. https://doi.org/10.1001/jama.2016.12534. PMID:27607056.
64. Cox D. Big Data and precision. *Biometrika*. 2015;102(3):712-716. https://doi.org/10.1093/biomet/asv033.
65. Filice CE, Joynt KE. Examining race and ethnicity information in Medicare administrative data. *Med Care*. 2016;Jul:29. PMID:27479593.
66. Kaneshiro B, Geling O, Gellert K, Millar L. The challenges of collecting data on race and ethnicity in a diverse, multiethnic state. *Hawaii Med J*. 2011;70(8):168-171. PMID:21886309.
67. Kass RE, Caffo BS, Davidian M, Meng XL, Yu B, Reid N. Ten Simple Rules for Effective Statistical Practice. *PLOS Comput Biol*. 2016;12(6):e1004961. https://doi.org/10.1371/journal.pcbi.1004961. PMID:27281180.
68. Hochster HS, Niedzwiecki D. Big Data, Small Effects. *J Clin Oncol*. 2016;34(11):1170-1171. https://doi.org/10.1200/JCO.2015.65.8161. PMID:26884573.
69. Srinivasan S, Moser RP, Willis G, et al. Small is essential: importance of subpopulation research in cancer control. *Am J Public Health*. 2015;105(S3)(suppl 3):S371-S373. https://doi.org/10.2105/AJPH.2014.302267. PMID:25905825.
70. Mardis ER. The challenges of big data. *Dis Model Mech*. 2016;9(5):483-485. https://doi.org/10.1242/dmm.025585. PMID:27147249.
71. National Institutes of Health. *NIH Data Sharing Policy and Implementation* Guidance: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm;2003. Accessed October 18, 2016.
72. Luo J, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomed Inform Insights*. 2016;8:1-10. https://doi.org/10.4137/BII.S31559. PMID:26843812.
73. Bahga A, Madisetti VK. A cloud-based approach for interoperable electronic health records (EHRs). IEEE J Biomed Health Inform. 2013;17(5):894-906. https://doi.org/10.1109/JBHI.2013.2257818. PMID:25055368.
74. Sobhy D, El-Sonbaty Y, Abou Elnasr M. MedCloud. Health care cloud computing system. 2012 International Conference for Internet Technology and Secured Transactions. 2012:161-166.
75. Lin WM, Dou WC, Zhou ZJ, Liu C. A cloud-based framework for Home-diagnosis service over big medical data. *J Syst Softw*. 2015;102:192-206. https://doi.org/10.1016/j.jss.2014.05.068.
76. Denny JC. Chapter 13: mining electronic health records in the genomics era. *PLOS Comput Biol*. 2012;8(12):e1002823. https://doi.org/10.1371/journal.pcbi.1002823. PMID:23300414.
77. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med*. 2010;153(9):600-606. https://doi.org/10.7326/0003-4819-153-9-201011020-00010. PMID:21041580.
78. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. *J Am Med Inform Assoc*. 2014;21(4):576-577. https://doi.org/10.1136/amiajnl-2014-002864. PMID:24821744.
79. Weber GM, Murphy SN, McMurry AJ, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc*. 2009;16(5):624-630. https://doi.org/10.1197/jamia.M3191. PMID:19567788.
80. Doshi JA, Hendrick FB, Graff JS, Stuart BC. Data, data everywhere, but access remains a big issue for researchers: a review of access policies for publicly-funded patient-level health care data in the United States. *EGEMS (Wash DC)*. 2016;4(2):1204. https://doi.org/10.13063/2327-9214.1204. PMID:27141517.
81. Warren E. Strengthening Research through Data Sharing. *N Engl J Med*. 2016;375(5):401-403. https://doi.org/10.1056/NEJMp1607282. PMID:27518656.
82. National Science Foundation NCfSaES. *Women, Minorities, and Persons with Disabilities in Science and Engineering.* Arlington, VA: Special Report NSF 15-

311. https://www.nsf.gov/statistics/2015/nsf15311/digest/;2015. Accessed October 18, 2016.
83. Valantine H A, Collins FS. National Institutes of Health addresses the science of diversity. P*roceedings of the National Academy of Sciences*. 2015;112(40):12240-12242. https://doi.org/10.1073/pnas.1515612112.
84. Van Horn JD. Opinion: big data biomedicine offers big higher education opportunities. Proc Natl Acad Sci USA. 2016;113(23):6322-6324. https://doi.org/10.1073/pnas.1607582113. PMID:27274038.
85. National Institute of Minority Health and Health Disparities. National Institutes of Health; http://grants.nih.gov/grants/guide/rfa-files/RFA-MD-16-002.html: 2016. Accessed October 18, 2016.
86. McEligot AJ, Behseta S, Cuajungco MP, Van Horn JD, Toga AW. Wrangling Big Data Through Diversity, Research Education and Partnerships. *Calif J Health Promot*. 2015;13(3):vi-ix. PMID:27257409