Arizona Legal Studies
Discussion Paper No. 16-26

Big Data Neglects Populations Most in Need of Medical and Public Health Research and Interventions

Sarah E. Malanga
Jonathan D. Loe
Christopher T. Robertson
Kenneth S. Ramos
The University of Arizona
James E. Rogers College of Law

August 2016

Malanga, S.E., Loe, J.D., Robertson, C.T., and Ramos, K.S. Big Data Neglects Populations Most in Need of Medical and Public Health Research and Interventions. In H.F. Lynch, I.G. Cohen, and U. Gasser (eds)., *Big Data, Health Law, and Bioethics*. Cambridge, United Kingdom: Cambridge University Press. Forthcoming 2017.

Big Data Neglects Populations Most in Need of Medical and Public Health Research and Interventions

Sarah E. Malanga, Jonathan D. Loe, Christopher T. Robertson, and Kenneth S. Ramos

INTRODUCTION

Scholars and commentators have expressed concern that "big data" may be getting too big, as too much of our lives is being tracked, captured, and analyzed, sometimes without our knowledge or permission. We here express a different concern: in one important way "big data" is not big enough.

Big data refers to a set of emerging technologies designed to "extract value from large volumes of a wide variety of data by enabling high-velocity capture, discovery, and analysis" (Villars, Olofson, & Eastwood, 2001). These technologies include pattern recognition, data repositories, and natural-language recognition (Costa, 2014). The information gleaned from big data can be extremely useful to those in the medical and public health fields engaged in research, intervention, program implementation, and behavioral analysis (Margolis et al., 2014). Furthermore, big data can be helpful in recruiting study participants and designing specific interventions. In the emerging field of precision medicine, for instance, more targeted, individualized treatment plans can be created by taking advantage of big data generated by modern "omics" technologies and the relatively easy integration of these data with the electronic health record ("EHR") (Costa, 2013).

The problem is that big data is largely homogenous, failing to reflect the growing diversity of the United States population. To date, big data has not captured certain marginalized

Electronic copy available at: http://ssrn.com/abstract=2828954

demographics. Particularly concerning are racial minorities (African-American, Hispanics, and Native American), people with low socioeconomic status, the LGBT community, and immigrants (Rabess, 2014; Lerman, 2013). Many of the people missing from the data that comes from sources such as internet history, social media presence, and credit card use are also missing from other sources of big data, such as EHRs and genomic databases. The factors responsible for these gaps are diverse and include lack of insurance, inability to access healthcare, and low levels of health literacy, to list a few, which leaves those missing from the data at an even greater disadvantage and more susceptible to missing out on the healthcare advantages and benefits that big data can provide.

Further exacerbating the problem is the fact that many of the people who are unable to integrate into the large data trail are the very people most in need of increased health research, intervention, and care. Compelling evidence increasingly supports the association between social determinants of health and health disparities among certain populations (Penman-Aguilar et al., 2016). Minority populations, for instance, can suffer higher burden of disease, which can be disproportionately concentrated in certain conditions, such as asthma, hypertension, and diabetes (Bodenheimer, Chen, & Bennett, 2009; Egede & Dagogo-Jack, 2005). Moreover, drug effectiveness varies among groups based on specific variations in their genomes (Piquette-Miller & Grant, 2007). Thus, using precision medicine to target populations disproportionately affected by health disparities is of utmost importance. However, in order to effectively utilize big data in conjunction with individualized medicine to tailor medical and public health research, programming, and interventions to specific populations, the necessary information must be available in the databases used to advance big data platforms. If big data can be made representative of the diversity of the US populations, we will be better able to realize the benefits

Electronic copy available at: http://ssrn.com/abstract=2828954

of precision medicine as evidenced by improvements in early diagnosis, development of targeted therapies with increased effectiveness and reduced toxicities, prevention of disease, and improved health outcomes.

This paper posits that the scope of big data and the subsequent studies and interventions designed around the information gathered needs to be broadened to increase the diversity of the data by capturing the heterogeneity of the population, not only in conventional demographic terms, but importantly, in terms of those most vulnerable to diminished health outcomes. Part one of this paper discusses big data and its value for medicine and public health, including behavioral policy, healthcare, and precision medicine. Part two discusses the failure of big data to adequately represent certain populations, which also suffer from the most severe health disparities – a problem that stymies the utility of big data. Finally, part three of this paper provides recommendations on how best to increase the amount of big data being collected on specific populations.

BIG DATA AND THE PROMISE FOR HEALTH

Traditionally defined by the three V's – volume, velocity, and variety – big data is the compilation, sorting, and organization of vast amounts of data (Andreu-Perez, Poon, Merrifield, Wong, & Yang, 2015). Many discussions of big data now also include three new V's – variability, veracity, and value (Andreu-Perez et al., 2015). Essentially, new and emerging technologies ease the processing of large amounts of data and make searching and categorizing the data to address specific needs much more feasible (Peters, 2014). With such ease comes the potential to use big data in a number of sectors – business, analytics, marketing, and medicine (Peters, 2014). In terms of the medical sector, data production is both structured and unstructured and the data collected can provide information about genomics, proteomics, metabolomics, and

the physiological features of a given individual, among other information (Andreu-Perez et al., 2015).

Big data is collected from a variety of diverse sources, including, but not limited to, credit card use information, social media profiles, and webpage visits. Information is also mined from many health related sources such as health insurance claims, EHRs, and data on pharmaceutical drug prescription and usage trends (Onukwugha, 2016). Another source for big data is behavioral and biometrics data derived from health applications on smartphones, wearables like FitBit and the Apple Watch, and everyday consumer products equipped with data tracking sensors and microchips, such as refrigerators and thermostats (Hofacker, Malthouse, & Sultan, 2016).

The data produced is being applied to the healthcare industry in a number of ways (Costa, 2014). One approach utilizes the data to spot trends that might otherwise go unnoticed. Google Trends, for instance, tracks diseases by analyzing spikes in Google searches (Costa, 2014). Search terms have allowed health officials and clinicians to respond to influenza, breast cancer, and bariatric surgery (Williams & Smith, 2015). And the FDA has piloted a program that combines surveillance, assessment, and simple, efficient probing of "diverse automated healthcare data holders – like electronic health record systems, administrative and insurances claims databases, and registries" to monitor the safety of FDA-regulated medical products including drugs, vaccines, and medical devices (FDA's Sentinel Initiative, 2016).

Big data can then help create and adapt interventions to meet the needs identified (Pentland, Reid, & Heibeck, 2013). For example, the National Drug Early Warning System monitors both social media and traditional data sources. The system uses the information to "detect emerging drug trends so that public health officials can launch community interventions to prevent the spread of illicit drug use" (Thorpe & Gray, 2015). The idea is to help communities

to respond effectively to emerging problems, as the Broward County community was able to do recently in the face of flakka, a synthetic drug (Frankel, 2016).

Big data is also valuable when applied to precision medicine as it allows more "targeted diagnostics and treatment based on each patient's history, ancestry and genetic profile" (Costa, 2014). Furthermore, the use of big data analytics will save time and improve the quality and effectiveness of healthcare delivery and patient care through the use of targeted therapies, informatics tools, and computer-aided diagnostics (Costa, 2014). Costa cites the case of Kalydeco, a new drug developed by Vertex Pharmaceuticals. The company screened hundreds of thousands of compound combinations using computer software and the help of more than 200 scientists to create a drug that treats some 4% of cystic fibrosis patients with the G551D mutation (Costa, 2014).

Many industries use big data to tailor their products to better serve the populations to which they cater. Manufacturers, for instance, use big data to adapt their products designs and features to meet specific needs. For example, a major biopharmaceuticals manufacturer increased a vaccine's yield by more than 50% using analytical tools to identify process parameters and their impact on yield and make changes accordingly (Auschitzky, Hammer, & Rajagopaul, 2014). Other companies, such as marketing and public relations firms, use big data to create more successful advertising campaigns. The medical community can use comprehensive data sources to inform the design of research studies as well as the creation and implementation of individualized interventions (Costa, 2014). The field of precision medicine uses big data to improve clinical trials, repurpose pharmaceuticals, optimize therapies, define the heterogeneity of chronic disease, and redefine the taxonomy of disease, among others (Jameson & Longo, 2015). Genome sequencing and genetic testing, for example, enables clinicians to identify

whether or not an individual may be at risk for specific cancers or conditions, such as hyperparathyroidism, and can screen patients accordingly while avoiding unnecessary screening in those who do not present a higher risk (Jameson & Longo, 2015).

One of the main elements of public health is the design and implementation of health interventions, such as smoking cessation counseling via "quitlines" and workplace programs aimed at decreasing obesity (Educational, 2016). Big data can be used to make such interventions more specific and relevant to the communities and populations being served. For example, data gleaned from grocery store receipts can allow tracking of nutrition for a population living in proximity to a store (Ransley et al., 2001). This sort of data can, in turn, lead to other programs or interventions being implemented in the community to continuously address and alleviate the health issues that may arise.

Big data has also been used globally to address issues of concern in the area of population genetics. Iceland, for instance, has been at the forefront of developing a large scale, population genetic database (Gibbons, Helgason, Kaye, Nomper, & Wendel, 2005). The database, run by deCODE, captures 18 years of genomic information for nearly a third of the Icelandic population (Palmer, 2015). The information gathered by deCODE has been used in conjunction with medical data from the country to identify disease related variants within the genomes (Palmer, 2015). The database was purchased by Amgen, a pharmaceutical company, to design targeted interventions within the population (Hirschler, 2012).

India has also entered the world of big data through an identification program aimed at making it easier for India's 1.2 billion citizens to prove their identity (Kumar, 2013). While the identification system was pioneered with the hopes of cutting back on government corruption, in the future the system will also be used in the health sector and will enable access to an

individual's health data (Kumar, 2013). Because the system is being used to deliver payments that help those in poverty and track previously unmonitored rural populations, the same populations that may be targeted for health interventions.

These are only a few examples. They illustrate that, around the world, big data holds big promises for healthcare.

BIG DATA EXCLUSIONS AND HEALTH DISPARITIES

Unfortunately, if big data is not sufficiently inclusive, these benefits may not extend to populations who are already underserved and marginalized by society – such as racial and ethnic minorities, people with low socioeconomic status, the LGBT community, and immigrants – the very same groups who are at risk for increased health disparities (Rabess, 2014). Table 1 summarizes the relative exclusion of these groups from big data.

Consumer data is the first major source of big data that may be exploited for health purposes. It comes from smartphone applications, wearable devices, and general internet use (e.g., patterns of health-related searching). Approximately 20% of Americans do not use internet in any form--at home, at work, at school, or by mobile phone (Wyatt, 2013). Moreover, home internet access reflects and reproduces deeper social and economic disparities. Only 57% of African-American households and 58% of Hispanic households use the internet, compared with 67% of white households (National Telecommunications, 2013). Education is important also--only 37% of households without a high school diploma have internet access at home, against 90% of households with a college degree (National Telecommunications, 2013). Older Americans have less access--just over half of Americans aged over 65, compared with over three-fourths for those under (National Telecommunications, 2013). Similar percentages apply to

those with (48%) and without (76%) a disability (National Telecommunications, 2013). And the overall rate of internet adoption has not improved since roughly 2009. (Wyatt, 2013).

Similarly, smartphone ownership (along with the health apps and integrated fitness tracking devices that come with them) reflect some disparities. Only 30% of Americans over the age of 65 own smartphones, as compared to 86% of those aged between 18 and 29, and 83% of 30-49 year olds (Anderson, 2015). Educational disparities continue too: 41% of those without a high school diploma own such a phone, while 81% of those with a college degree do (Anderson, 2015). Low-income populations do not have access to smartphones at the same rate as richer Americans: 52% of those making less than $30,000 as against 87% for those making more than $75,000. Finally, rural populations own phones at a 52% rate, compared to 72% of those in urban areas (Anderson, 2015). Nonetheless, blacks, whites, and Hispanic American adults own smartphones at similar rates (Anderson, 2015).

Statistics for wearables are not as clear. Overall, the ownership rate for wearables is lower than that of smartphones (only 10% of American adults own a fitness tracker (NPD, 2015). It does seem likely that we can expect their uptake to continue to reflect this trend. For example, 48% of fitness tracker owners have an income above $100,000, a far higher number than the general population.

The second major source of big data is derived from EHRs and health insurance claims. As technology has improved, the ability to store health related information electronically has become increasingly easier. This information includes health history, doctors and emergency room visits, prescription and treatment history, and insurance claims. In 2009, as part of the American Recovery and Reinvestment Act, the Health Information Technology for Economic and Clinical Health (HITECH) was passed to "advance the implementation of a nationwide

health IT infrastructure that improves healthcare quality, reduces health disparities, and advances delivery of patient-centered medical care, among other goals" (NORC, 2010). Included in the HITECH Act is the Medicare and Medicaid Electronic Health Record Incentive Programs, which provides funding to professionals and hospitals that meaningfully use EHR technology (NORC, 2010). The Act also emphasizes the use of health technology to treat underserved populations, providing technical assistance and EHR implementation at Federally Qualified Health Centers and favorable incentives to Medicaid providers (NORC, 2010).

As of 2014, 97% of all reported non-federal acute care hospitals had certified EHR technology and 74.1% of all office-based physicians had a certified EHR system (Charles, Gabriel, & Searcy, 2015; Jamoom, Yang, & Hing, 2016). The data suggest disparities between medical specialties (e.g., cardiologists and family/general practitioners had greater uptake, while psychiatrists had lower uptake), but use of EHR is not shown to be associated with patient demographics (Kokkonnen et al., 2013). Nonetheless, even if an individual has access to their EHR, there is no guarantee they understand the information provided (NORC, 2010). Moreover, even if the use of EHRs is widespread amongst physicians, 53% of patients say they are unable to access their health records, either because they do not have internet access or do not know how to do so, and, therefore, will not receive the potential benefits of EHRs, such as improved communication between patients and physicians (Heath, 2015). Minorities are also significantly less likely to enroll in online patient portals, which are useful for both accessing EHR's as well as providing additional patient interactions and data (Goel et al., 2011).

Although the use of EHRs continues to expand, they of course only cover persons who have sought healthcare, a predicate which itself suffers from disparities due to lack of health insurance coverage. Minorities continue to have higher uninsured rates. In 2014, 11.8% of

Blacks, 9.3% of Asians, and 19.9% of Hispanics were uninsured as compared to 7.6% of non-Hispanic Whites (Smith & Medalla, 2015). People living in poverty also experienced much higher uninsured rates. Those living below 100 percent poverty had the highest uninsured rate (19.3%) while those living at 400 percent or above poverty had the lowest rate (4.8%) (Smith & Medalla, 2015). Uninsured rates across all levels of poverty were higher in those states that did not expand Medicaid eligibility than in those states that did (Smith & Medalla, 2015). Older members of the population had higher insured rates in comparison to younger members of the population as did people with higher levels of education (Smith & Medalla, 2015). Uninsured individuals are less likely to access healthcare (Coey, 2015), and even if they do so, will not appear in health insurance claims databases.

Another problem that arises when collecting data from these specific populations is the distrust that many members of these populations have towards the medical community. This distrust, especially among minority populations, stems from instances of exploitation and unethical treatment at the hands of medical professionals including the Havasupai diabetes project and the Tuskegee experiment on African American men as well as an overall fear of racism and dissimilar treatment (Pacheco et al., 2013; Kennedy, Mathis, & Woods, 2007). These incidents have led to a wariness among minorities and a lack of participation in research studies, organ donation, and other healthcare initiatives that are beneficial to their health (Pacheco et al., 2013; Russell, 2012). This history may not only deter people from participating in research but may also deter them from seeking medical care altogether. So deterred, this only amplifies the disparate representation in a wide range of data sources, including EHRs.

As mentioned briefly above, these exclusions reflect pre-existing health disparities, and may amplify them. The medical, public health, and scientific literature emphasize the many

factors that can influence an individual's health outcomes. Some of these factors can potentially be controlled, such as diet for instance, while others, such as genetics, less so. In this regard, the emerging field of epigenetics will transform the interface between genetics, lifestyle, and environment given the remarkable reversibility of epigenetics changes in the regulation of genetic and phenotypic expression. The unequal dissemination of socioeconomic status, race, ethnicity, and education leads to unequal health outcomes among certain sectors of the population (Brondolo, Gallo, & Myers, 2009). Racial and ethnic minorities in the US are at greater risk for a number of diseases and health issues, including heart disease and hypertension, diabetes, and poor birth outcomes (Brondolo et al., 2009). Health disparities include not only heightened levels of morbidity and mortality but also an unequal degree of access to health care, health insurance, and other factors that impact health outcomes (Health Disparities, 2011). All of these factors are indicators of poor health outcomes among both children and adults, and these health disparities are magnified when an individual falls into more than one of these "disadvantaged" categories (Braveman, Cubbin, Egerter, Williams, & Pamuk, 2010).

It is apparent that these disadvantaged categories are the same as those groups with poorer representation in the sources of big data. Given their poorer outlook as it stands, it is troubling that the promise of big data for health improvement discussed in Section I may not fully be shared and the gap in health outcomes potentially increased. In particular, the absence of applicable data makes it nearly impossible to apply precision medicine to these populations.

The heightened risk of health disparity among disadvantaged populations can lead to poor health outcomes for individuals, their families, and their communities, putting entire subpopulations at risk. In order to decrease the disproportionate effect these factors have on select populations, policy change, interventions, and research are needed. The health issues

themselves must be addressed as well as the underlying causes including poor nutrition, inadequate health care, lack of infrastructure, unsuitable programming, and outright neglect, to list a few.

Unfortunately, when available data on a given population is skewed, interventions and research studies aimed at addressing the needs of those specific populations will not be successful. We may generate biased estimates as to the likely effects of interventions and distortions of the research agenda -- from deciding which medical conditions to research to the development of precise drugs and devices to address them. Such distortions may have epistemic consequences, if they affect the recruitment of study participants, for example.

These problems may be most trenchant in the domain of precision medicine, which promises to individualize care using research and clinical practice (Mirnezami, Nicholson, & Darzi, 2012). This field is harnessing diagnostic, prognostic, and therapeutic strategies precisely tailored to each patient's requirements (Mohler, Najafi, & Fain, 2015), but depends on a robust dataset that allows easier classification of diseases, sequencing genomes, and the making of diagnoses based on imaging (Jameson & Longo, 2015). The proliferation of big data influences precision medicine in that it can provide information on everything from prescription drug use to popular dietary trends to prevalence of smoking within a certain community. However this precision cannot occur without the necessary data to inform diagnosis and treatment.

Although much of the discussion on big data tends to focus on issues of privacy and civil liberties, it is important to note that the exclusivity of big data is troublesome because of its potential ramifications in terms of national health, both presently and in the future. Neglecting to gather information about minority populations, those who live below the poverty line, and other marginalized sectors of society, does not simply mean there is a dearth of information about

these populations, it also means that these populations, who are most at risk for a number of health disparities, including increased heart disease, higher rates of infant mortality, and domestic violence, will not benefit from the tailored attention, care, and benefits that big data can yield.

RECOMMENDATIONS

Given the documented absence of vulnerable minorities from sources of big data, and the profoundly positive effect on the individuals and populations who are recipients of big data benefits, there is a strong incentive to develop solutions to this absence--to make big data inclusive enough. Although addressing and correcting deficits within the world of big data is an effort that will require collaborations among the federal government, medical professionals, and industry, here we document some efforts to do so, and propose others. Table 2 summarizes these recommendations.

We should acknowledge that for some purposes these problems can be solved through standard post-hoc statistical adjustments that yield unbiased estimates notwithstanding known disparities in the way the data is collected. For example, statisticians using EHR's to track population health have acknowledged that the data is likely to suffer from a selection bias on an income gradient (NYC Health, 2013). However, patient-level income is not likely to appear in the EHR data, making simple stratification techniques unavailable. "The most feasible approach may be to standardize the crude estimates to the age, sex, race, and income distribution of the population to which the data are to be generalized (defined as poststratification adjustment), using neighborhood-level income as a proxy" (NYC Health, 2013 at 11).

During his 2015 State of the Union Address, President Obama announced the Precision Medicine Initiative (Office of the Press Secretary, 2015). Enabled with $215 million in funding,

the Initiative aims to encourage investment and research into precision medicine and to provide medical professionals with the appropriate tools and knowledge to integrate precision medicine into clinical practice (Dennis, 2015). The main objectives of the initiative include increasing and improving cancer treatment, the creation of a national research cohort that will include the voluntarily contributed data of one million Americans, modernizing the current regulatory scheme, and collaboration between public and private sectors (Office of the Press Secretary, 2015). We applaud the goal expressed by the Precision Medicine Initiative that the proposed research cohort will "broadly reflect the diversity of the U.S. population" and will include participants from different social and ethnic groups (About). While the new funds made available through the Initiative in February of this year are currently focused on building the cohort, additional resources could be leveraged to increase the number of grants focused on diverse populations in the future.

The increased focus on precision medicine by NIH grew out of a number of factors. These factors not only include the initiative but also the growing ability to sequence the human genome, improvements in analytics technologies, and new tools that enable the use of large datasets as well as enthusiasm from the medical community. NIH has rapidly responded to the call of the initiative. In March of 2015, for example, a team of experts in precision medicine and clinical research was formed to seek public input, define a vision for cohort participation, and outline what can be gained from the study as well as what obstacles might arise (NIH News Releases, 2015). Funding was also provided to Vanderbilt University to begin a pilot project focused on the cohort (Neergaard, 2016). These steps are important advances, but it is critical that they are not limited to specific groups and in fact reach across the diverse segments of the population.

The Food and Drug Administration ("FDA") may be a powerful ally in helping to alleviate these unintended downfalls of the incomplete use of big data. The FDA has broad discretion on how and what it enforces. In its approval of certain pharmaceuticals and other health related products the FDA could require that the data used to inform the efficacy and safety of these items come from a diverse, inclusive pool. Currently the National Institutes' of Health ("NIH") Policy on Inclusion of Women and Minorities in Clinical Research only applies to research funded by NIH (Office of Women's Health, 2011). The goal of the policy, which was mandated by Congress in 1993 following a number of iterations that grew out of the women's health movement, is to guarantee that women and minorities are included in human subject research and requires inclusion "in numbers adequate to allow for valid analyses of difference in intervention effect" (Background, 2015).

The FDA's own recommendations, on the other hand, do not require the inclusion of minorities in industry-sponsored trials, but rather suggest the use of "a standardized approach for collecting and reporting race and ethnicity information in clinical trials conducted in the United States and abroad for certain FDA regulated products" (Guidance for Industry, 2005). In 1997 the Administration enacted its Modernization Act, which, among other issues, addressed the inclusion of women and minorities in clinical trials (FDAMA, 1998). The agency's "Dialogues on Diversifying Clinical Trials," expanded on this theme, providing a number of strategies that could increase the involvement of minorities in clinical trials including community outreach, trial re-design, and more efficient regulation (Office of Women's Health, 2011). While these guidelines provide information on how to uniformly collect the racial and ethnic data of study participants, they do not "address the level of participation of racial and ethnic groups in clinical

trials" (Guidance for Industry, 2005). The FDA should follow NIH's lead to require diverse data to inform the safety and efficacy of these drugs and devices.

Additionally, the FDA is increasingly relying on "omics" data to carry out its regulatory mandate (Torti, 2013). These efforts have now been framed within the context of guidelines that have been open for public comment. The report outlines the initiatives and efforts undertaken by the agency to ensure that its scientific base is robust, effective, and targeted to its regulatory responsibility to protect and promote health through ensuring the safety and effectiveness of human and veterinary drugs, biologics, and devices and the safety of foods and cosmetics. The scientific strategy builds upon five principles designed to specify implementation plans, deliverables and timelines, development of a preemptive approach, enhancement of infrastructure and core expertise in modern technologies, and effective communication. Several reports have been published to address these priorities (FDA Science, 2007*;* Future of Drug Safety, 2006; Challenges, 2007). In parallel, other federal agencies such as the National Cancer Institute have developed a checklist criteria for determination of the readiness of 'omics-based tests for guiding patient care in clinical trials (McShane et al., 2013). The criteria cover issues related to specimens, assays, modeling, design, and ethical, legal and regulatory aspects. As such, more efforts should be made to increase participation of stakeholders in vetting and updating these recommendations.

The FDA could also leverage its enforcement discretion over a number of mobile apps, such as those that provide assistance with smoking cessation or help track medication usage (Examples, 2015). While the FDA recognizes that many of these apps may meet the definition of a medical device and may be intended for use in diagnosis, treatment, or prevention, the Administration has decided, for now, not to subject such apps to regulatory requirements

(Examples, 2015). It could draw ideas from its "Dialogues on Diversifying Clinical Trials," in which the FDA lists a number of strategies that could increase the involvement of minorities in clinical trials including community outreach, trial re-design, and more efficient regulation (Office of Women's Health, 2011). In the mobile apps space, the FDA could require compliance with some of these same suggestions in an effort to diversify the pool from which data is mined. Beyond the safety and efficacy mandate that applies to drugs and devices, however, it is not clear that the FDA has such authority to require diversification of data collection for other purposes.

Another remedy may be for the federal government to issue new regulations aimed at this very issue in particular. Currently, much of the legal discussion around big data has centered around privacy and other issues of autonomy. For instance, any personal data collected is usually preceded by collection of informed consent from the individual to use their data (Perera, Ranjan, Wang, Khan, & Zomaya, 2015). In addition, data often times is anonymized in order to protect privacy when both collecting and disseminating the information gathered from individuals, especially if the information may be considered to be sensitive (Perera et al., 2015). Much of the dialogue is also concentrated on an individual's control over the information they produce and how the information should be managed by the individuals creating it (Perera et al., 2015).

Perhaps what is needed to address this specific issue and curb what Lerman identifies as the "reinforcement of the status of already disadvantaged groups" is a regulatory scheme that guarantees that, at the very least, any data being used to inform healthcare initiatives, development, and delivery is inclusive (Lerman, 2013). Because big data comes from so many unique sources, it may not be realistic or even possible to regulate *how* big data is collected, but it may be possible to regulate the ways in which data is searched, organized, and ultimately used (Mantelero & Vaciago, 2015).

Regulating the way in which big data is mined and used could face a number of barriers, and it is not a remedy that could be immediately put in place. Nevertheless, regulation has the potential to provide safeguards in terms of who is benefitting from the collection of data. Increased regulation targeting the mining of data may force those using the data to include information from more diverse populations, sources, and geographic locations – helping to guarantee that those who leave a smaller data trail are still included in the overall analysis and application of the data.

One model for solving the problem of collecting big data on certain populations is the federal Lifeline program. Introduced in 1985 as part of the Federal Communications Commission, it provides discounted, affordable phone services to low-income consumers (Lifeline, 2016). The program aims to give consumers the "opportunities and security" that come with telephone access (Lifeline, 2016). Lifeline was expanded in 1996 when it was made available to consumers across the country and again in 2005 when wireless services were included in the program (Lifeline, 2016; Honig, 2013). Additionally, in 2016 the FCC announced that broadband services would be added to the program, giving low-income Americans the opportunity for "full and meaningful participation in society" (Office of Media Relations, 2016). The adoption of cell phones into the program as well as the proposal to expand services to include broadband has the potential to increase the data produced by populations that may not otherwise have access to or engage with technology on a regular basis. The expansion of the program also suggests that this specific problem with regard to big data demographics may be remedied over time as more and more people own cell phones and have greater internet access (Anderson, 2015). It is important to note, however, that the services provided by Lifeline are limited, which, in turn, may limit the scope of data derived from these specific sources.

CONCLUSION

Big data is proving to be extremely useful in the medical and public health fields. It can provide degrees of information on business practices, attitudes, views and medical needs at both the individual and group levels that have not been previously available. As such, the benefits of big data and its application to research, pharmaceutical innovation, intervention design, and treatment plans are significant. Regrettably, not everyone is able to receive the benefits that big data provides because information is not currently being gathered in ways that ensure adequate representation of all segments of society. This is particularly significant in the U.S. given the "melting pot" nature of our diverse populations.

In order to address and remedy this oversight, big data collection must be augmented to include diverse populations. While this will not eradicate the health disparities that plague disadvantaged populations, the use of big data to design and implement more tailored medicine will be better poised to benefit the populations most at risk for health disparities. These precise studies, interventions, and treatments have the potential to positively impact the health outcomes of these populations both over the life course of the individuals as well as among future generations.

REFERENCES

About the Precision Medicine Initiative Cohort Program. *National Institutes of Health*. Retrieved from https://www.nih.gov/precision-medicine-initiative-cohort-program

Anderson, M. (2015) Technology Device Ownership: 2015. *Pew Research Center*. Retrieved from http://www.pewinternet.org/files/2015/10/PI_2015-10-29_device-ownership_FINAL.pdf

Andreu-Perez, J., Poon, C.C.Y., Merrifield, R.D., Wong, S.T.C., & Yang, G-Z. (2015). Big Data for Health. *IEEE J Biomed Health Inform*, *19*(4), 1193-1208.

Auschitzky, E., Hammer, M., & Rajagopaul, A. (2014). How big data can improve manufacturing. *McKinsey Company*. Retrieved from http://www.mckinsey.com/business-functions/operations/our-insights/how-big-data-can-improve-manufacturing

Background. (2015). *National Institutes of Health*. Retrieved from http://orwh.od.nih.gov/research/inclusion/background.asp

Bodenheimer, T., Chen, E., & Bennett, H.D. (2009). Confronting the Growing Burden of Chronic Disease: Can the U.S. Health Care Workforce Do the Job? *Health Affairs*, *28*(1), 64-74.

Braveman, P.A, Cubbin, C., Egerter, S., Williams, D.R., & Pamuk, E. (2010). Socioeconomic Disparities in Health in the United States: What the Patterns Tell Us. *J Public Health*, *100*(S1), 186-196.

Brondolo, E., Gallo, L.C., & Myers, H.F. (2009) Race, racism, and health: disparities, mechanisms, and interventions. *J Behav Med*, *32*, 1-8.

Challenges for the FDA. (2007). Challenges for the FDA: The Future of Drug Safety. *Institute of Medicine*.

Charles, D., Gabriel, M, & Searcy, T. (2015). Adoption of Electronic Health Record Systems among U.S. NonFederal Acute Care Hospitals: 2008-2014. *The Office of the National Coordinator for Health Information Technology*. Retrieved from https://www.healthit.gov/sites/default/files/data-brief/2014HospitalAdoptionDataBrief.pdf

Coey, D. (2015). The Effect of Medicaid on Health Care Consumption of Young Adults. *Health Econ*. *24*(5), 558-565.

Costa, F.F. (2014) Big Data in Biomedicine. *Drug Discov Today*, *19*(4), 433-440.

Costa, F.F. (2013) Social networks, web-based tools and diseases: implications for biomedical research. *Drug Discov Today*, *18*, 272-281.

Dennis, B. (2015, Jan. 29). Obama touts 'lifesaving' potential of personalized medicine. *The Washington Post*. Retrieved from https://www.washingtonpost.com/national/health-science/obama-seeks-215-million-for-personalized-medicine-initiative-using-genetic-data/2015/01/29/75789342-a7f4-11e4-a7c2-03d37af98440_story.html

Educational and Community-Based Programs. (2016). Educational and Community-Based Programs: Evidence-Based Resources. *Healthy People*. Retrieved from https://www.healthypeople.gov/2020/topics-objectives/topic/educational-and-community-based-programs/ebrs

Egede, L.E., & Dagogo-Jack, S. (2005). Epidemiology of Type 2 Diabetes: Focus on Ethnic Minorities. *MEd Clin N Am*. *89*, 949-975.

Examples of Mobile Apps for Which the FDA Will Exercise Enforcement Discretion (2015, September 22). *U.S. Food and Drug Administration*. Retrieved from http://www.fda.gov/MedicalDevices/DigitalHealth/MobileMedicalApplications/ucm368744.htm

FDA Action Plan to Enhance the Collection and Availability of Demographic Subgroup Data. (2014). *U.S. Food and Drug Administration*. Retrieved from http://www.fda.gov/downloads/RegulatoryInformation/Legislation/FederalFoodDrugandCosmeticActFDCAct/SignificantAmendmentstotheFDCAct/FDASIA/UCM410474.pdf

FDA Science and Mission at Risk. (2007). FDA Science and Mission at Risk: Report of the Subcommittee on Science and Technology. *U.S. Food and Drug Administration*. Retrieved from http://www.fda.gov/ohrms/dockets/ac/07/briefing/2007-4329b_02_01_FDA%20Report%20on%20Science%20and%20Technology.pdf

FDA's Sentinel Initiative. (2016, March 2). *U.S. Food and Drug Administration*. Retrieved from https://docs.google.com/document/d/1WWcPlQd5ypu2yPI551pYptq8Er-vnSCP1ynVIS-FEO0/edit

FDAMA. (1998). FDAMA Women and Minorities Working Group Project. *U.S. Food and Drug Administration*. Retrieved from http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm080616.pdf

Frankel, T.C. (2016, April 4). The surprising disappearance of flakka, the synthetic drug that pushed South Florida to the brink. *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/wonk/wp/2016/04/04/the-mysterious-disappearance-of-flakka-the-synthetic-drug-that-pushed-south-florida-to-the-brink/

Future of Drug Safety. (2006). The Future of Drug Safety: Promoting and Protecting the Health of the Public. *Institute of Medicine*.

Gibbons, S.M.C., Helgason, H.H., Kaye, J., Nomper, A., & Wendel, L. (2005). Lessons from European Population Genetic Databases: Comparing the Law in Estonia, Iceland, Sweden and the United Kingdom. *Eur J Health Law, 12*, 103-133.

Goel, M.S., Brown, T.L., Williams, A., Hasnain-Wynia, R., Thompson, J.A., & Baker, D.W. (2011). Disparities in Enrollment and Use of an Electronic Patient Portal. *J Gen Intern Med*. *26*(10), 1112-1116.

Guidance for Industry: Collection of Race and Ethnicity Data in Clinical Trials. (2005). *U.S. Food and Drug Administration*.

Heath, S. (2015, Dec. 31). 53% of Consumers Can't Access Electronic Health Record Info. *EHR Intelligence*. Retrieved from https://ehrintelligence.com/news/53-of-consumers-cant-access-electronic-health-record-info

Health Disparities and Inequalities Report – United States, 2011. (2011). *Centers for Disease Control and Prevention*.

Hirschler, B. (2012, Dec. 10). Amgen buys Icelandic gene hunter Decode for $415 million. *Reuters*. Retrieved from http://www.reuters.com/article/us-amgen-decode-idUSBRE8B90IU20121210

Hofacker, C.F., Malthouse, E.C., & Sultan, F. (2016). Big Data and consumer behavior: imminent opportunities. *J Consum Mark*, *33*(2), 89-97.

Honig, D. (2013, May 13). The Truth About Lifeline. *The Huffington Post*. Retrieved from http://www.huffingtonpost.com/david-honig/the-truth-about-lifeline_b_3266143.html

Jameson, J.L., & Longo, D.L. (2015). Precision Medicine – Personalized, Problematic, and Promising. *N Engl J Med*, *372*(23), 2229-2234.

Jamoom, E. W., Yang, N., & Hing, E. (2016). Adoption of Certified Electronic Health Record Systems and Electronic Information Sharing in Physician Offices: United States, 2013 and 2014. *U.S. Department of Health and Human Services*. Retrieved from http://www.cdc.gov/nchs/data/databriefs/db236.pdf

Kennedy, B.R., Mathis, C.C., & Woods, A.K. (2007). African Americans and Their Distrust of the Health Care System: Healthcare for Diverse Populations. *J Cult Divers*. *14*(2), 56-60.

Kokkonen, E.W.J., Davis, S.A., Lin, H.C., Dabade, T.S., Feldman, S.R., & Fleischer, A.B. (2013). Use of electronic medical records differs by specialty and office settings. *J Am Med Inform Assoc*. *20*(1), e33-e38.

Kumar, H. (2013, May 24). Unique ID Program Introduces Instant Verification Services. *The New York Times*. Retrieved from http://india.blogs.nytimes.com/2013/05/24/aadhar-program-introduces-instant-verification-services/?_r=0

Lerman, J. (2013). Big Data and Its Exclusions. *Stan L Rev Online*, 66, 55-63. Retrieved from http://www.stanfordlawreview.org/online/privacy-and-big-data/big-data-and-its-exclusions

Lifeline Support for Affordable Communications. (2016). *Federal Communications Commission*. Retrieved from https://www.fcc.gov/consumers/guides/lifeline-support-affordable-communications

Mantelero, A., & Vaciago, G. (2015). Data protection in a big data society. Ideas for a future regulation. *Digital Investigation*, 15, 104-109.

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J.,…Green, E.D. (2014) The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc*, *21*, 957-958.

McShane, L.M., Cavenagh, M.M., Lively, T.G., Eberhard, D.A., Bigbee, W.L., William, P.M.,...Conley, B.A. (2013). Criteria for the use of omics-based predictors in clinical trials. *Nature*. *502*, 317-320.

Mirnezami, R., Nicholson, J., & Darzi, A. (2012) Preparing for Precision Medicine. *N Engl J Med*, *366*(6), 489-491.

Mohler, J., Najafi, B., & Fain, M. (2015). Precision Medicine: A Wider Definition. *J Am Geriatr Soc*, *63*(9), 1971-1972.

National Telecommunications and Information Administration & Economics and Statistics Administration. (2013). Exploring the Digital Nation: America's Emerging Online Experience. *Department of Commerce*. Retrieved from www.ntia.doc.gov/files/ntia/publications/exploring_the_digital_nation_-_americas_emerging_online_experience.pdf

Neergaard, L. (2016, Feb. 15). Retreived from http://www.pbs.org/newshour/rundown/nih-taking-first-steps-to-huge-precision-medicine-project/

NIH News Releases. (2015, Mar. 30). NIH taking first steps to huge precision medicine project. *Associated Press*. Retrieved from http://www.nih.gov/news-events/news-releases/nih-forms-team-experts-chart-course-presidents-precision-medicine-initiative-research-network

NORC. (2010). Understanding the Impact of Health IT in Underserved Communities and those with Health Disparities. *Norc at the University of Chicago*.

NPD. (2015). The Demographic Divide: Fitness Trackers and Smartwatches Attracting Very Different Segments of the Market, According to The NPD Group. Retrieved from https://www.npd.com/wps/portal/npd/us/news/press-releases/2015/the-demographic-divide-fitness-trackers-and-smartwatches-attracting-very-different-segments-of-the-market-according-to-the-npd-group/

NYC Health. (2013). Developing an Electronic Health Record-Based Population Health Surveillance System. *New York City Department of Health and Mental Hygiene*. Retrieved from http://www.nyc.gov/html/doh/downloads/pdf/data/nyc-macro-report.pdf

Office of Media Relations. (2016). FCC Modernizes Lifeline Program for the Digital Age: New Rules Will Help Make Broadband More Affordable for Low-Income Americans. *Federal Communications Commission*. Retrieved from https://apps.fcc.gov/edocs_public/attachmatch/DOC-338676A1.pdf

Office of the Press Secretary. (2015). *FACT SHEET: President Obama's Precision Medicine Initiative.* Retrieved from https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative

Office of Women's Health. Dialogues on Diversifying Clinical Trials: Successful Strategies for Engaging Women and Minorities in Clinical Trials. (2011). *Food and Drug Administration Office of Women's Health*.

Onukwugha, E. (2016). Big Data and Its Role in Health Economics and Outcomes Research: A Collection of Perspectives on Data Sources, Measurement, and Analysis. *Pharmaco Economics*, *34*, 91-93.

Pacheco, C.M., Daley, S.M., Brown, T., Filippi, M., Greiner, A., & Daley, C.M. (2013). Moving Forward: Breaking the Cycle of Mistrust Between American Indians and Researchers. *Am J Public Health*. *103*(12), 2152-2159.

Palmer, K.M. (2015, March 25). Why Iceland is the World's Greatest Genetic Laboratory. *Wired*. Retrieved from http://www.wired.com/2015/03/iceland-worlds-greatest-genetic-laboratory/

Penman-Aguilar, A., Talih, M., Huang, D., Mooneshinge, R., Bouye, K., & Beckles, G. (2016). Measurement of Health Disparities, Health Inequities, and Social Determinants of Health to Support the Advancement of Health Equity. *J Public Health Manag Pract*, *22*(1 Supp), S33-S42.

Pentland, A., Reid, T.G., & Heibeck, T. (2013). Revolutionizing Medicine and Public Health. *World Innovation Summit for Health Big Data and Health Working Group*.

Perera, C., Ranjan, R., Wang, L., Khan, S.U., & Zomaya, A.Y. (2015) Big Data Privacy in the Internet of Things Era. *IT Pro*, *17*(3), 32-39.

Peters, S.G., & Buntrock, J.D. (2014). Big Data and the Electronic Health Record. *J Ambulatory Care Manage*, *37*(3), 206-210.

Piquette-Miller, M., and Grant, D.M. (2007). The Art and Science of Personalized Medicine. *Clin Pharmacol Ther*, *81*(3), 311-315.

Rabess, C.E. (2014). Can Big Data Be Racist? Retrieved from http://www.thebolditalic.com/articles/4502-can-big-data-be-racist

Ransley, J.K, Donnelly, J.K, Khara, T.N., Botham, H., Arnot, H., Greenwood, D.C., & Cade, J.E. (2001).
The use of supermarket till receipts to determine fat and energy intake in a UK population. *Public Health Nutr*. *4*(6), 1279-1286.

Russell, E., Robinson, D.H.Z., Thompson, N.J., Perryman, J.P., & Arriola, K.R.J. (2012). Distrust in the Healthcare System and Organ Donation Intentions Among African Americans. *J Community Health*. *37*, 40-47.

Smith, J.C., & Medalla, C. (2015). Health Insurance Coverage in the United States: 2014. *United States Census Bureau*.

Thorpe, J.H., & Gray, E.A. (2015) Big Data and Public Health: Navigating Privacy Laws to Maximize Potential. *Public Health Reports*, *130*, 171-175.

Torti, F.M. (2013). Report on Status of Regulatory Science at FDA: Progress, Plans and Challenges. *U.S. Food and Drug Administration*. Retrieved from http://www.fda.gov/AdvisoryCommittees/CommitteesMeetingMaterials/ScienceBoardtotheFoodandDrugAdministration/ucm115716.htm

Villars, R.L., Olofson, C.W., & Eastwood, M. (2001). Big Data: What It Is and Why You Should Care. *IDC*.

Williams, R.F., & Smith, G.P. (2015). Using "Big Data" to Optimize Public Health Outreach. *JAMA Dermatol*, *151*(4), 367-368.

Wyatt, E. (2013, August 18). Most of U.S. Is Wired, but Millions Aren't Plugged In. *N.Y. Times*. http://www.nytimes.com/2013/08/19/technology/a-push-to-connect-millions-who-live-offline-to-the-internet.html?_r=0

Table 1

Big Data source	How it's Used	Gaps in Collection
Social Media (i.e. Facebook, Twitter); Wearables (i.e., Apple Watch, FitBit); Internet Use	Consumer trends, behavioral data, and biometrics	Not everyone has access to social media, wearables, or the internet; not everyone chooses to engage with social media or use a wearable.
Electronic Health Records; Insurance Claims	Gathering and analyzing data on a range of medical and healthcare issues – disease prevalence, diagnoses, prescription drug trends	Certain hospitals/healthcare facilities may not use EMRs, therefore creating a dearth of information on certain demographics
Genome Sequencing; Biospecimens; Research Participation	Ability to collect and analyze genetic information	Cost prohibitive; emphasis on collecting from certain populations

Table 2

Big Data source	What's Currently Being Done to Address Those Gaps?	Recommendation
Social Media (i.e. Facebook, Twitter); Wearables (i.e., Apple Watch, FitBit); Internet Use	Lifeline was expanded in March 2016 to include broadband services, which will be available to low-income Americans.	Expand Lifeline program to provide smartphones and mobile internet.
Electronic Health Records; Insurance Claims	HITECH Initiative; ACA; Medicaid expansion	Utilize EHRs in a more diverse, representative manner – expand their use across the country, with focus on populations that are currently not receiving adequate representation
Genome Sequencing; Biospecimens; Research participation	NIH policy; FDA guidelines	FDA should make diversity guidelines mandatory and enforce data standards in the research it will accept in the approval process.