Commentary: In the Rising Era of Big Data, Small Steps are Key

Nina Heller[1]; Jonathan H. Seltzer, MD, MBA, MA[1]

Big Data may be useful to identify and perhaps ameliorate health disparities. However, over reliance on the power on Big Data can potentially cause harm. When assessing health disparities, the use of Big Data should be limited to hypothesis generation.
Ethn Dis. 2017;27(2):73-76; doi:10.18865/ed.27.2.73.

Keywords: Big Data, Health Disparities

[1]ACI Clinical, Bala Cynwyd, Pensylvania

Address correspondence to Jonathan H. Seltzer, MD; ACI Clinical; 251 St. Asaphs Road, #402; Bala Cynwyd, PA 19004; 484.429.7200; jseltzer@aciclinical.com

Commentary

A few years ago, some very smart consultants from McKinsey & Company wrote about how the Big Data revolution will accelerate value and innovation in health care.[1] Indeed, government, health care institutions and medical product companies have invested heavily in Big Data. Are they planning to revolutionize the disparities in health care? There are not too many revolutions in any area where those at society's bottom rung benefit. We believe that scientific progress does not come in revolutions, but rather one step forward at a time and the advent of Big Data should be viewed from that perspective.

When it comes to evaluating public health issues, we rely on data analysis. Recently, the use of Big Data has been gaining currency as the answer for health care decision makers. The pharmaceutical industry and health care community have gathered lots of Big Data and herald its promise for health care decision-making. Proponents champion the power of Big Data and claim it equals the efficacy of clinical trial data at a fraction of the cost. In theory, large datasets show promise for improving health care decision-making at both population and subpopulation levels. The hope is that Big Data will help us to better understand and aid in bridging the health gap related to income, education level, sex, race, ethnicity, employment status, and sexual orientation. However, before we fully embrace Big Data, we need to understand whether it is ready for public exposure and if its use will decrease or increase disparities within our health care system.

…before we fully embrace Big Data, we need to understand whether it is ready for public exposure and if its use will decrease or increase disparities within our health care system.

First, let's look at how health care decisions are currently informed

without the Big Data bandwagon in tow. Most everyone will agree that the gold standard for health care decision-making is the clinical trial. Generally, clinical trials are designed to evaluate whether an intervention, for example a drug or medical device, actually works. Clinical trials have been a key driver in forming the technological success of modern medicine over the past 40 years. Additionally, large clinical trials demonstrated potential disparity in cancer, heart disease and diabetes care and have revealed that different treatments may benefit different populations.[2-6] Clinical trials have served as excellent sources for decision making because they meet three best practices recommendations: 1) rigorous, pre-defined rules for asking a question (eg, does this treatment lower blood pressure); 2) data collection that is specific to the question being asked (eg, specify how and when to collect blood pressure measurement, laboratories, medication use); and 3) data quality is of the highest integrity (eg, quality control processes, which confirm the completeness and accuracy of the data).

Unless these best practices are applied properly, conclusions from clinical trial data can be manipulated. An example where best practices are not followed is in some of the treatment guidelines from The American Heart Association/American College of Cardiology (AHA/ACC). The AHA/ACC guidelines retrospectively 'combined' numerous clinical trials to create guidelines that recommend optimal treatment strategies.[7-9] (Violation of best practice #1: not pre-specifying the question asked prior to data collection). As a result, these treatment guidelines do not account for the known differences in myocardial infarction and atrial fibrillation seen in African Americans and women.[10,11] Several colleagues and I examined the clinical trial data that support the AHA/ACC guidelines and found that the supporting data are only in small part derived from minority and female patients.[7] Despite these known differences, the AHA/ACC guidelines do not take into account health disparities. Thus, due to not using best practices, health disparities can easily be ignored. Even worse, payers and regulators have adopted these standards as "quality care" thereby perpetuating health disparities.

…it is imperative that the adoption of Big Data is approached with caution.

This situation may be amplified when it comes to Big Data. In the above example, we see what happens when only one best practice is ignored. Big Data often ignores all three. Unlike clinical trials, Big Data is collected before the data's purpose is established. Currently, most Big Data is collected to facilitate financial or operational goals, not to improve health care. For example, claims databases and electronic health records were designed for financial and operational purposes and accordingly were designed to answer questions like how many lab tests and CT scans were performed. Big Data was not designed to answer questions about whether the lab tests were appropriate or whether the CT scan was necessary. Furthermore, perhaps more importantly, they are not designed to tell us if the diagnosis was correct, or if different groups react differently to treatments or doctors' instructions.

Additionally, when it comes to data integrity, there is no infrastructure to support the integrity of Big Data. The enormous size of data that must be managed when accruing massive datasets is cause for problems. In particular, patient-level data are now collected across several platforms such as medical devices, prescriptions, self-reported data, and personal devices (fitness trackers, apps, and health-related internet searches).[12] This massive data collection becomes more vulnerable to overlooking gaps within the data during analysis, in particular minorities who may not have access to several of these platforms.

That said, there is enormous power in Big Data to support statistically based decision making. Good statistical practices require transparency about assumptions supported by strong methodologies. At the current time, we are not convinced that either of these two conditions are consistently applied. In their absence, the reliance upon statistical models can result in a high prevalence of confirmation bias resulting in the distortion and misrepresentation of data. Therefore, if Big Data, as currently constituted, is used to examine the

many issues involved with health disparities, there is large potential for the manipulation of Big Data to fit whatever mold one wishes.

The accessibility of Big Data is quite appealing to many researchers as it provides a vast amount of information collected in a quick and inexpensive manner. The enormous amount of information, however, and if not handled properly-- either through incompetence or bias-- can lead to spurious or divergent conclusions as we've recently seen with the novel anticoagulant agents.[13] Therefore, it is imperative that the adoption of Big Data is approached with caution. Chai and Shih recently published a set of guidelines for using Big Data effectively: 1) beware of spurious correlations in open-ended searches; 2) be conscious of sample sizes and sample variation when mining for correlations; 3) beware of systematic biases in data collection. These guidelines provide a framework for developing analytical pipelines by researchers without falling into the pitfalls mentioned above. The authors remind the scientific community that despite the progressive milestones Big Data holds, not to forget traditional scientific methods.[14] Big Data collected with these guidelines in place and used in conjunction with the traditional scientific methods of experimentation, theoretical models, computer modeling, and simulation may be the key to preventing perpetuation of health disparity and conclusions researchers are sifting through datasets so hard find.

Big Data is not useless, but at the current time it should be limited in helping policy makers create and examine sources and hypotheses about health disparities. Testing these hypotheses will likely take a combination of research approaches such as the improvement of current Big Data solutions to include datasets that are specifically created to examine health disparities. Additionally, it is likely that clinical trial levels of evidence may be necessary to find these solutions. We are at the start of the Big Data revolution; we should move forward carefully.

Conflict of Interest

No conflicts of interest to report.

Author Contributions

Research concept and design: Seltzer; Acquisition of data: Seltzer, Heller; Data analysis and interpretation: Heller; Manuscript draft: Seltzer, Heller; Administrative: Heller; Supervision: Seltzer

References

1. Kayyali B, Knott D, Van Kuiken S. The big-data revolution in US health care: Acceleration value and innovation. McKinsey & Company. http://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care. Published April 2013. Accessed November 15, 2016.
2. Denise Grady. In cancer trials, minorities face extra hurdles. *The New York Times*. https://www.nytimes.com/2016/12/23/health/cancer-trials-immunotherapy.html?_r=0. Published December 23, 2016. Accessed February 10, 2017.
3. White RO, Beech BM, Miller S. Health care disparities and diabetes care: practical considerations for primary care providers. *Clin Diabetes*. 2009;27(3):105-112. https://doi.org/10.2337/diaclin.27.3.105. PMID:21289869.
4. Herman WH, Young MA, Uwaifo G et. al. Differences in A1C by race and ethnicity among patients with impaired glucose tolerance in the Diabetes Prevention Program. *Diabetes Care*. 2007;30(10): 2453–2457.
5. Potosky AL, Harlan LC, Kaplan RS, Johnson KA, Lynch CF. Age, sex, and racial differences in the use of standard adjuvant therapy for colorectal cancer. *J Clin Oncol*. 2002;20(5):1192-1202. https://doi.org/10.1200/JCO.20.5.1192. PMID:11870160.
6. Soliman EZ, Safford MM, Muntner P, et al. Atrial fibrillation and the risk of myocardial infarction. *JAMA Intern Med*. 2014;174(1):107-114. https://doi.org/10.1001/jamainternmed.2013.11912. PMID:24190540.
7. Wann LS, Curtis AB, January CT, and the ACCF/AHA Task Force on Practice Guidelines. 2011 ACCF/AHA/HRS focused update on the management of patients with atrial fibrillation (updating the 2006 guideline). *Circulation*. 2011;123(1):104-123. https://doi: 10.1161/CIR.0b013e3181fa3cf4.
8. Wright RS, Anderson JL, Adams CD, et al. 2011 ACCF/AHA focused update of the Guidelines for the Management of Patients with Unstable Angina/Non-ST-Elevation Myocardial Infarction (updating the 2007 guideline): a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines developed in collaboration with the American College of Emergency Physicians, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons. *J Am Coll Cardiol*. 2011;57(19):1920-1959. https://doi.org/10.1016/j.jacc.2011.02.009. PMID:21450428.
9. Hunt SA, Abraham WT, Chin MH, et al; American College of Cardiology Foundation; American Heart Association. 2009 Focused update incorporated into the ACC/AHA 2005 Guidelines for the Diagnosis and Management of Heart Failure in Adults A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines Developed in Collaboration With the International Society for Heart and Lung Transplantation. *J Am Coll Cardiol*. 2009;53(15):e1-e90. https://doi.org/10.1016/j.jacc.2008.11.013. PMID:19358937.
10. Sardar MR, Badri M, Prince CT, Seltzer J, Kowey PR. Underrepresentation of women, elderly patients, and racial minorities in the randomized trials used for cardiovascular guidelines. *JAMA Intern Med*. 2014;174(11):1868-1870. https://doi.org/10.1001/jamainternmed.2014.4758. PMID:25264856.
11. Barrett MA, Humblet O, Hiatt RA, Adler NE. Big Data and disease prevention: From quantified self to quantified communities. *Big Data*. 2013;1(3):168-175. https://doi.org/10.1089/big.2013.0027. PMID:27442198.
12. Taleb NN. Beware the big errors of 'Big Data.' Wired.com. 2013. Available at: http://www.wired.com/2013/02/big-data-means-big-errors-people/. Published Febru-

ary 8, 2013. Accessed November 15, 2016.
13. Perry S. Newest blood-thinning drugs are the subject of a troubling, in-depth series on their safety. MinnPost; 2015. Available at: https://www.minnpost.com/second-opinion/2015/08/newest-blood-thinning-drugs-are-subject-troubling-depth-series-their-safety. Accessed February 10, 2017.
14. Chai S, Shih W. "Why Big Data Isn't Enough." *MIT Sloan Manag Rev*. 2017;58(2):57-61. Available at: http://sloan-review.mit.edu/article/why-big-data-isnt-enough/. Accessed February 10, 2017.