American Journal of Epidemiology

Vol. 179, No. 9

Book Review

Big Data: A Revolution That Will Transform How We Live, Work, and Think

By Viktor Mayer-Schönberger and Kenneth Cukier

ISBN-10: 0544227751, ISBN-13/EAN: 9780544227750, Houghton Mifflin Harcourt, Boston, Massachusetts (Telephone: 855-969-4642, Fax: 800-269-5232, E-mail: myhmhco@hmhco.com, Website: http://www.hmhco.com/), 2013, 256 pp., $27.00 Hardcover, $12.76 Paperback, $15.95 eBook

"Big data" have big potential to influence society and science, including the way we approach and conduct epidemiologic research. In this book, Mayer-Schönberger and Cukier (1) outline 3 perspectives (or "shifts") which are (the authors argue) inherent in big data: 1) the obsolescence of sampling (chapter 2, "More"); 2) the acceptance of increased measurement error in return for more data (chapter 3, "Messy"); and 3) a "move away from the age-old search for causality" (chapter 4, "Correlation") (1, p. 14).

The book is divided into 10 chapters. Chapter 5 discusses the move in industry and government to collect data on everything (including health-related topics); chapter 6 discusses the value of such data. Chapters 7–9 discuss the implications and risks of this practice and ways to control the "datafication" of human life. Chapter 10 summarizes future possibilities. In this review, we focus on chapters 2–4, which explore the 3 shifts which the authors claim big data will bring to the scientific table and which we believe will be of most interest (and most concern) to readers of the *Journal*.

The chapter entitled "More" (chapter 2) explains how big data is defined by collecting as much data as possible, and at times "all" of it. The authors argue that new technologies will allow data scientists to passively collect, store, and analyze much more data in real time. In many instances, the authors refer to sampling as an outdated hindrance to discovery: "Reaching for a random sample in the age of big data is like clutching at a horse whip in the era of the motor car. . . . Increasingly, we will aim to go for it all" (1, p. 31). However, the authors seem to misunderstand random sampling. Describing an analysis in which "all the data" were used (in this case, every sumo match over 11 years) to discover a surprising pattern of match-throwing, the authors claim that "random sampling of the bouts might have failed to reveal it" because "without knowing what to look for, one would have no idea what sample to use" (1, p. 29). As far as we can understand with regard to this passage, the authors are confusing issues of precision and validity: A simple random sample of the 11 years' worth of sumo bouts might have had less *power* to reveal the finding than an analysis of all bouts, but it is nonsense to suggest that the same pattern would not have emerged in a simple random sample of the bouts.

Of course, if it is possible and straightforward to do so, using all of the data is ideal; yet the perspective taken by the authors ignores the fact that a sample is defined relative to a target population (2), and we may not be able to collect data on "n = all" because of how we define our target. Relatedly, the authors fail to appreciate that their examples of "all the data" are frequently no such thing. Google Flu Trends (Google, Inc., Mountain View, California)—a frequent example in these chapters—uses a great deal of data, but it is data obtained from a subset of literate and (most likely) high-socioeconomic-status individuals. If we were interested in a disease that was more concentrated among persons without access to computers, would Google searches work as well? Similar criticisms can be made of the authors' claim that

> when the data is [sic] collected passively while people do what they normally do anyway, the old biases associated with sampling . . . disappear. We can now collect information that we couldn't before, be it relationships revealed via mobile phone calls or sentiments unveiled through tweets (1, p. 30).

The idea that information revealed through tweets is free from bias is simply risible: Twitter (Twitter Inc., San Francisco, California) is used by perhaps 16% of Americans (3). Moreover, public health science is frequently faced with finite resources: It is far from clear that it is in the interests of public health to devote resources to (for example) expensive assays in a very large sample of individuals when a simple random sample will provide the same point estimate and negligibly worse precision.

Chapter 3 ("Messy") discusses measurement error, an inherent part of any data science and a central concern in epidemiology (4). The authors claim that an embrace of big data allows for greater acceptance of measurement error, and argue that "[i]n return for relaxing the standards of allowable errors, one can get ahold of much more data, [and] sometimes 'more trumps better' " (1, p. 33). While the authors parenthetically mention a caveat about the need to avoid "systematic bias" (1, p. 34), they fail to explain that merely having greater amounts of systematically biased data will often lead to more problems rather than fewer. A meta-analysis of 10 studies with the same measurement error bias will be more precise, but no more valid, than any single one of those studies. At the same time, the increased precision will offer increased (but false) confidence in those fundamentally flawed results. It is not until much later in the book that the authors acknowledge that big data can exacerbate the issue of "relying on the numbers when they are far more fallible than we think" (1, p. 163). To expand on this: Whether a larger but messier data set is preferable to a smaller but less messy data set is a substantive question on which we believe epidemiologists should take a consequentialist view (5): The data set to be

preferred is the one that will yield an answer that will in turn lead to the greatest improvement in human health. It is far from obvious that more is always better—indeed, more may sometimes be less.

Chapter 4, entitled "Correlation," brings some clarity on the motivation behind statements made in the previous chapters. It also reveals a pivotal fallacy in the authors' thinking about the general utility of "big data" as defined here. They predict that, in the future, "big-data correlations will routinely be used to disprove our causal intuitions, showing that often there is little if any statistical connection between the effect and its supposed cause" (1, p. 64). Thus, the authors argue, if we believe that E causes D, but E is not associated with D in a correlational analysis of big data, we can safely conclude that E does not cause D. Of course, this is wrong: Confounding bias (among other types of bias) can just as easily erase a true causal effect as create a spurious one.

Throughout this chapter, the authors offer examples of how estimation of correlations in big data has yielded insights leading to improved practices—for instance, predicting infection in premature infants in intensive care 24 hours before overt symptoms appear (6) or predicting who will be readmitted to the hospital after discharge (1, p. 128), without regard for causal mechanism. To their credit, regarding the latter example, the authors state outright that the prediction of readmission "says nothing to establish causality . . . [but] nevertheless suggests that a post-discharge intervention" might reduce readmissions and costs (1, p. 128). The authors are right that such predictive analyses can play an important role in improving health, even without establishing causality; but at the same time, they fail to appreciate the central role of understanding causality in improving public health, and too often conflate predictive and etiologic inference.

For example, the authors imagine that if "millions of electronic medical records reveal that cancer sufferers who take a combination of aspirin and orange juice see their disease go into remission, the exact cause for the improvement in health may be less important than the fact they lived" (1, p. 14). Their phrasing raises the question of the underlying goal of the exercise. If our goal is merely to identify patients likely to go into remission in the future, then the authors are correct; but if our goal is to help additional people who are suffering with cancer go into remission, then it matters a great deal whether it is a combination of orange juice and aspirin that *caused* the remissions or rather some other, associated factor. If it is the latter, obviously, a recommendation to drink more orange juice and take more aspirin is unlikely to have an effect on individual or population health. We recall early (perhaps apocryphal) notions that scurvy was associated with not being on land, and therefore sailors were treated by being buried neck-deep in dirt, to no avail. The authors' confusion is underlined later when they claim that "big data does not tell us anything about causality" (1, p. 163), in contrast to their earlier (erroneous) claim that lack of correlation can disprove causality, and in apparent disregard for the ways we can (with effort and assumptions) tease out causal inferences from observational data.

A major theme in this book is that "big data" will become the dominant scientific paradigm, and (in this and other ways) change society—and it may yet. Chapters 5–10 do a good and highly readable job of giving examples of how big-data approaches may transform both business and health in the coming decades, as well as the potential dangers of the datafication of society. However, the perspective of this book offhandedly discounts decades of work in numerous fields with little justification or explanation, in a way rife with misconceptions. We agree with the authors that science and public health are at the cusp of a major and important change, in which "big data" will play an integral role. Yet it seems equally clear that the perspectives offered in this book would benefit from a firmer grounding in existing scientific approaches and perspectives, and thus at present they may have relatively little utility for the practicing epidemiologist.

ACKNOWLEDGMENTS

Dr. Ashley I. Naimi was supported by a postdoctoral research fellowship from the Fonds de recherche du Québec-Santé.

Conflict of interest: none declared.

REFERENCES

1. Mayer-Schönberger V, Cukier K. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Houghton Mifflin Harcourt; 2013.
2. Moore DS, McCabe GP. *Introduction to the Practice of Statistics*. New York, NY: W.H. Freeman and Company; 2012.
3. Fung B. Not that many Americans use Twitter, apparently. *The Washington Post*. November 4, 2013. (http://www.washingtonpost.com/blogs/the-switch/wp/2013/11/04/nobody-uses-twitter-in-america-apparently/). (Accessed February 3, 2014).
4. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *Int J Epidemiol*. 2006;35(4):1074–1081.
5. Galea S. An argument for a consequentialist epidemiology. *Am J Epidemiol*. 2013;178(8):1185–1191.
6. Blount M, Ebling MR, Eklund JM, et al. Real-time analysis for intensive care: development and deployment of the Artemis analytic system. *IEEE Eng Med Biol Mag*. 2010;29(2):110–118.

Ashley I. Naimi[1] and Daniel J. Westreich[2]
(e-mail: ashley.naimi@mcgill.ca)
[1] *Department of Obstetrics and Gynecology, Faculty of Medicine, McGill University, Montreal, Quebec H3A1A1, Canada*
[2] *Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599*

DOI: 10.1093/aje/kwu085; Advance Access publication: April 8, 2014