Epidemiology in the Era of Big Data

Stephen J. Mooney,[a] *Daniel J. Westreich,*[b] *and Abdulrahman M. El-Sayed*[a]

Abstract: Big Data has increasingly been promoted as a revolutionary development in the future of science, including epidemiology. However, the definition and implications of Big Data for epidemiology remain unclear. We here provide a working definition of Big Data predicated on the so-called "three V's": variety, volume, and velocity. From this definition, we argue that Big Data has evolutionary and revolutionary implications for identifying and intervening on the determinants of population health. We suggest that as more sources of diverse data become publicly available, the ability to combine and refine these data to yield valid answers to epidemiologic questions will be invaluable. We conclude that while epidemiology as practiced today will continue to be practiced in the Big Data future, a component of our field's future value lies in integrating subject matter knowledge with increased technical savvy. Our training programs and our visions for future public health interventions should reflect this future.

The popular and scholarly press has—with considerable excitement—begun using the term "Big Data" to describe the rapid integration and analysis of large-scale information.[1–3] However, a clear definition of Big Data remains elusive, and the ways by which Big Data's advent might shape the future of epidemiologic research and population health intervention remain unclear.[4] Although previous authors have considered the role of Big Data in clinical care,[2,5–7] we are herein concerned with its implications for the future of research and practice of epidemiology and population health.

BIG DATA: WHAT IS IT?

The characterization of Big Data has evolved since the term was coined in the computer science literature in 1997 to refer to data too large to be stored in then-conventional storage systems.[8] One increasingly accepted[7] designation revolves around the "3V's": *high-variety*, *high-volume*, and/or *high-velocity* information assets.[9] Under this definition, *high variety* refers to the practice of incorporating data collected originally for disparate purposes into a single dataset for combined analysis, such as combining data from electronic medical records with purchase histories or social media profile updates.[3] *High volume* refers to data with orders of magnitude more observations and/or orders of magnitude more variables per observation than prior datasets in the domain. And *high velocity* refers to a data generation process wherein data are compiled and analyzed in real-time or nearly in real-time, often by algorithms operating without human intervention.

Submitted 24 October 2014; accepted 27 January 2015.

From the [a]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY; and [b]Department of Epidemiology, Gillings School of Public Health, University of North Carolina, Chapel Hill, NC.

S.J.M. was supported by the National Cancer Institute at the National Institutes of Health (T32-CA09529). D.J.W. was partially supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number DP2HD084070. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Correspondence: Stephen J. Mooney, Department of Epidemiology, Mailman School of Public Health, 722 West 168th Street, New York, NY 10032. E-mail: sjm2186@columbia.edu.

ISSN: 1044-3983/15/2603-0390
DOI: 10.1097/EDE.0000000000000274

Big Data in Historical Context

Modern technology has increased the quantity and forms of data available, and the speed with which they can be merged. While some aspects are new, integration and analysis of large-scale data has always been fundamental to epidemiology. John Graunt's analysis of weekly counts of burials in London parish cemeteries, circa 1662, was arguably the first ever epidemiologic analysis. Graunt's work represented a revolutionary "Big Data" approach at a time when the understanding of disease as the product of imbalance of 'humors' rendered inference from group comparison largely inconceivable.[37,38] Graunt's work compiled causes of death from approximately 3,500 sheets of paper into a single table nearly a century before the rubber eraser was invented, creating a novel, high-volume dataset with which to investigate causes of mortality. In working with records that were originally collected to provide early warning of plague outbreaks,[37,38] rather than to explore alternate causes of death, Graunt confronted issues germane to modern high-variety data. For example, his findings may have been biased by misclassification of the cause of death owing to reporting inaccuracies or by mass burials obscuring characteristics of individual deaths. Although Graunt demonstrated conclusively that plague outbreaks did not result from astrological or political events—commonly held theories of the time—it would take several centuries and the advent of germ theory to finally determine the etiology of the plague. About 200 years later, William Farr picked up Graunt's torch by organizing and developing death certificates in mid-19th-century United Kingdom. Farr's work, inspired by the need to ensure validity in analyses of mortality data, also increased the ease of constructing high-volume mortality datasets. Without this development, many early public health analyses, such as comparisons of mortality rates by profession over vast numbers of individuals, would have been unreliable or impossible.[39] About a hundred years after Farr's work, Hammond and Horn took advantage of an increasingly interconnected society to assemble a high-volume cohortfollowing nearly 188,000 subjects,[40] a large cohort even by today's standards. Hammond and Horn's work was enabled in part by new technologies such as punch-card computing,but was also built on a classic data source: death certificates. In our own time, Big Data has been touted as the future of science in general and epidemiology in particular.[2,24] But as we consider how Big Data may represent the future of our field, we should recall that it is also our past.

THE THREE V'S AND EPIDEMIOLOGY

High-Variety Data and Measurement Error

Within epidemiology, variety in data is not new (see text box above), having long been achieved by merging separately collected datasets. In some analyses, high-variety datasets are assembled from datasets collected independently but intended for epidemiologic inquiry, such as adding genomic data to survey responses, or adding environmental data in a gene–environment interaction study. In other examples, data are repurposed from repositories of data collected initially for other aims, such as New York City's OpenData initiative.[10] As administrative data are increasingly made available online, the bureaucratic challenge of merging such datasets is decreasing.

Although the increased quantity of data sources presents new opportunities, working with secondary data reinforces existing validity challenges. Epidemiologists have established that biases due to measurement error are independent of the volume of data.[11] However, some in the popular press have argued that the sheer quantity of information available in the age of Big Data may allow us to accept lower quality data.[2] In this context, it may be important for epidemiologists to influence the data gathering process to improve the validity of administratively collected data. Efforts to use low-quality data almost invariably result in calls for relevant data to be recorded accurately[7,12]—a strong argument for the involvement of epidemiologists at the design stages of administrative data collection systems in an era in which almost any data could be fruitfully repurposed for epidemiologic analyses.

High-Volume Data and Analytic Rigor

In addition to increasing the need for rigorous measurement, the increase in the variety of data described above will also lead to an increase in data volume, as more variables per subject create wider datasets. For example, genomic single nucleotide polymorphism microarrays can add thousands of columns per subject to a dataset.[13] Similarly, there are potentially hundreds of ways to define neighborhoods using geographic information systems and US Census data, each articulating different characteristics of social spaces, and so each adding a column to the width of the dataset.[14]

One response to the challenge of increasing dataset width is to use tools that aid with variable selection. Analyses testing causal hypotheses may require software to assist with developing directed acyclic graphs representing theorized data relations (eg, DAGitty).[15] Data explorations may use machine learning tools and other emerging technologies for so-called hypothesis-generating analyses.[16]

Technological innovations will likely also enable inclusion of more subjects in studies, resulting in taller datasets. Web-based and cellular technologies already enable much cheaper recruitment and follow-up of subjects than can telephone-based surveys.[17] Furthermore, as laboratory techniques develop and assay costs decline, molecular epidemiologists can enroll more subjects at the same cost,[13] and as integration of health systems continues, national-scale electronic health records studies will become more detailed and powerful.[5]

Increasing data width may require increased engagement with statistical and computational techniques, whereas increasing height may require increased engagement with underlying theory and subject matter knowledge to interpret results. It has been long recognized that substantive (or background) knowledge is necessary for etiologic inference,[18,19] but the need to distinguish between a highly precise finding and a finding with potential clinical or interventional importance will increase with population size.[20,21] With a sufficiently large analytic population, many statistical interaction terms will be accompanied by low *P* values, but this does not imply that such information can be used productively to improve population health.[22]

High-Velocity Data and Intervention Optimization

Instantaneous data collection holds promise for public health improvement, even if the rapidity with which data can be automatically collected or analyzed is not integral to all epidemiologic analysis. Several existing applications use high-velocity data for surveillance. For example, Google Flu Trends, which uses data from geo-located web searches to track influenza activity,[23] has served as an exemplar of a Big Data approach to surveillance, although with caveats.[24] Similarly, researchers have tracked other outcomes using Google search trends[25] and developed related systems to track the flu using Twitter.[26]

Increased data velocity may also be valuable for implementing interventions. This potential is particularly true where interventions must be deployed quickly in response to unfolding threats to population health and where information is the rate-limiting factor in optimizing such interventions. For example, the introduction of cholera to Haiti after the 2010 earthquake required a major public health response under adverse conditions.[27] Identification of infected subjects and deployment of available oral cholera vaccines would have been aided by the use of high-velocity technologies such as cellular networks. In practice, unfortunately, no vaccine was deployed in the early stages of the outbreak due to the difficulty of identifying the optimal population to vaccinate.[28]

High data velocity may also enable interventions to be designed with the intent of rapid iteration. For example, a program designed to enhance medication adherence might deploy pill dispensers equipped with technology to report, via the Internet or cellular networks, whether pills were dispensed on schedule.[29] Program developers could use this real-time technology to test different messaging strategies, using data from these pill dispensers as outcomes. Such interventions, which may also be available to any program using social media to effect behavioral change,[30] are analogous to the A/B testing frameworks that have enabled improvement to website user experiences through rapid experimentation.[31] These experiments, in which users are randomly assigned to one of two web experiences to determine effects of design changes on engagement metrics such as click-throughs or time spent at the site, may become valuable as public health messaging moves to web-based platforms. Of course, A/B testing must be applied only with sufficient attention to public health and research ethics.[32]

IMPLICATIONS OF BIG DATA FOR TRAINING

The Big Data future will require some epidemiologists to embrace technological skills not traditionally within the epidemiology portfolio, particularly computer programming. For example, with moderate programming skills and the required permissions, analytic datasets can be assembled from publicly available information using web-scraping programs that read and compile data from web pages.[33] Similarly, public health interventions designed for rapid iteration may need to leverage mobile applications or centralized servers to control and optimize interventions. A secondary benefit might be to broaden the pathways by which trained epidemiologists can improve population health. Many technology entrepreneurs build companies to encourage healthy lifestyles (eg, Noom, RunKeeper, MyFitnessPal) and in the process accumulate large repositories of behavioral data. Epidemiologists with the skills to engage directly with large-scale data and the methods to analyze it may find opportunities to collaborate with such enterprises for both academic and industry benefit.

We caution, however, that any training in software engineering must not come at the cost of training in core epidemiologic skills. For example, an analysis intended to determine regional variation in stigma due to sexual identity using Twitter would benefit from a principal investigator with skills to acquire the data from Twitter directly. However, it is more important that such an investigator be able to judge the value of tools to measure expressions of stigmatizing views, to formulate an analysis accounting for the fact that American users of Twitter are unlikely to represent Americans as a whole, and so on. Given the already large amount of material covered by graduate programs in epidemiology, computer programming may represent a specialized track of epidemiologic training for those who already have substantial expertise in a health-related domain. Increased recruitment of epidemiology graduate students from technical fields, whose undergraduates rarely enter epidemiology today, including computer science, may also help to increase the prevalence of these increasingly valuable skills among epidemiologists.

IMPLICATIONS OF BIG DATA FOR PRACTICE

Epidemiology's metric for success, including any value realized from Big Data, should be measured in terms of improvements in population health.[34] In the future, metrics may be gathered most efficiently using high-velocity technologies. The study of high-velocity feedback may then become a core component of the emerging field of implementation science.[35] For example, before A/B testing can be widely used in messaging-based interventions, best practices for its deployment in population health should be developed and validated.

By contrast, although epidemiologic practice will benefit from access to higher volume and higher variety data, such access is unlikely to revolutionize epidemiologic practice in the ways that some optimists have suggested,[2,36] such as obviating the need for causal theory, or eliminating classical challenges to validity associated with imperfect data. Therefore, the core of epidemiologic practice, that is, understanding the causes of population health and optimizing interventions to improve it, will remain conceptually and practically challenging in the Big Data era.

TABLE. Summary of the Three V's of Big Data and Their Implications

Name	Meaning	Examples	Opportunities and Challenges	Implications for Epidemiology and Public Health
Volume	Datasets with more observations	National electronic health record databases, social media datasets	Power to precisely measure unexpected associations, though potentially without substantive relevance	Evolutionary/incremental
Variety	Datasets with variables from different sources; more variables per observation	-omics data, neighborhood data added to a phone survey	Capacity to assess complex interactions, but more complicated variable selection	Evolutionary/incremental
Velocity	Data collected and analyzed in real time	Medication adherence intervention messaging adapted to subject response pattern	Potential to design dynamic interventions	Potentially revolutionary

CONCLUSIONS

Big Data holds promise to identify population health intervention targets through analysis of high-volume and high-variety data, and to target and refine ensuing interventions using high-velocity feedback mechanisms (Table). An agenda leveraging Big Data's potential would be best led by epidemiologists with skill sets rooted in traditional principles, and who are also comfortable with emerging technologies.

Tall, wide, and messy data are already available, but at present such data represent a trickle: now is the time to prepare for the oncoming flood. Although epidemiology as practiced today will continue to be practiced in a Big Data future, a component of our field's future value lies in integrating subject matter knowledge with increased technical savvy. Our training programs and our visions for future public health interventions should reflect that.

ACKNOWLEDGMENTS

Dr. Alfredo Morabia, Dr. Catherine Williams, and Dr. Sharon Schwartz gave insightful comments on an earlier version of this work.

REFERENCES

1. The New York Times. *Big Data Compendium*. Available at: http://www.nytimes.com/compendium/collections/576/big_data Accessed September 5, 2014.
2. Mayer-Schönberger V, Cukier K. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Houghton Mifflin Harcourt; 2013.
3. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. *JAMA*. 2014;311:2479–2480.
4. Fallik D. For big data, big questions remain. *Health Aff (Millwood)*. 2014;33:1111–1114.
5. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA*. 2013;309:1351–1352.
6. Bollier D, Firestone CM. *The Promise and Peril of Big Data*. Washington, DC: Aspen Institute, Communications and Society Program; 2010.
7. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. *Health Aff (Millwood)*. 2014;33:1115–1122.
8. Cox M, Ellsworth D. *Application-controlled demand paging for out-of-core visualization. Proceedings of the 8th Conference on Visualization'97*. IEEE Computer Society Press; 1997;235-ff.
9. Douglas L. *The Importance of "Big Data": A Definition*. Stamford, CT: Gartner; 2012.
10. The City of New York. *NYC Open Data*. Available at: http://www.nyc.gov/html/data/about.html. Accessed August 8, 2014.
11. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*. 1977;105:488–495.
12. Halamka JD. Early experiences with big data at an academic medical center. *Health Aff (Millwood)*. 2014;33:1132–1138.
13. Khoury MJ, Lam TK, Ioannidis JP, et al. Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiol Biomarkers Prev*. 2013;22:508–516.
14. Krieger N, Zierler S, Hogan JW, et al. Geocoding and measurement of neighborhood socioeconomic position: a US perspective. In: Kawachi I, Berkman L, eds. *Neighborhoods and Health*. Oxford, England: Oxford University Press; 2003:147–178.
15. Textor J, Hardt J, Knüppel S. DAGitty: a graphical tool for analyzing causal diagrams. *Epidemiology*. 2011;22:745.
16. Glymour MM, Osypuk TL, Rehkopf DH. Invited commentary: Off-roading with social epidemiology–exploration, causation, translation. *Am J Epidemiol*. 2013;178:858–863.
17. Cook C, Heath F, Thompson RL. A meta-analysis of response rates in web-or internet-based surveys. *Educ Psychol Meas*. 2000;60:821–836.
18. Krieger N. Epidemiology and the web of causation: has anyone seen the spider? *Soc Sci Med*. 1994;39:887–903.
19. Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology*. 2001;12:313–320.
20. Poole C. Low P-values or narrow confidence intervals: which are more durable? *Epidemiology*. 2001;12:291–294.
21. Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of P values and evidence. *J Am Stat Assoc*. 1987;82:112–122.
22. Siontis GC, Ioannidis JP. Risk factors and interventions with statistically significant tiny effects. *Int J Epidemiol*. 2011;40:1292–1307.
23. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin Infect Dis*. 2009;49:1557–1564.
24. Lazer DM, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. *Science*. 2014;343:1203–1205.
25. Seifter A, Schwarzwalder A, Geis K, Aucott J. The utility of "Google Trends" for epidemiological research: Lyme disease as an example. *Geospat Health*. 2010;4:135–137.
26. Lampos V, Cristianini N. *Tracking the flu pandemic by monitoring the social web*. 2nd International Workshop on Cognitive Information Processing (CIP). New York: IEEE Press; 2010:411–416.
27. Frerichs RR, Keim PS, Barrais R, Piarroux R. Nepalese origin of cholera epidemic in Haiti. *Clin Microbiol Infect*. 2012;18:E158–E163.
28. Date KA, Vicari A, Hyde TB, et al. Considerations for oral cholera vaccine use during outbreak after earthquake in Haiti, 2010− 2011. *Emerg Infect Dis*. 2011;17:2105.

29. Sutton S, Kinmonth A-L, Hardeman W, et al. Does electronic monitoring influence adherence to medication? Randomized controlled trial of measurement reactivity. *Annals Behav Med*. 2014:48:293–299.
30. Laranjo L, Arguel A, Neves AL, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. *J Am Med Inform Assoc*. 2015;22:243–256.
31. Kohavi R, Henne RM, Sommerfield D. *Practical guide to controlled experiments on the web: listen to your customers not to the hippo*. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: ACM; 2007:959–967.
32. Emanuel EJ, Grady CC, Crouch RA, Lie RK, Miller FG, Wendler DD. *The Oxford Textbook of Clinical Research Ethics*. Oxford, England: Oxford University Press; 2011.
33. Lee BK. Epidemiologic research and Web 2.0: the user-driven Web. *Epidemiology*. 2010;21:760–763.
34. Galea S. An argument for a consequentialist epidemiology. *Am J Epidemiol*. 2013;178:1185–1191.
35. El-Sadr WM, Philip NM, Justman J. Letting HIV transform academia–embracing implementation science. *N Engl J Med*. 2014;370:1679–1681.
36. Anderson C. The end of theory. *Wired Mag*. 2008;16:16–17.