Two Ways of Knowing: Big Data and Evidence-Based Medicine

Ida Sim, MD, PhD

Evidence-based medicine (EBM) is more than 20 years old (1). Although EBM's painstaking path of careful clinical studies, critical appraisal of published evidence, and methodologically rigorous systematic reviews has been the template for knowing what works in medicine, new "big data" approaches seem to offer a powerful and tempting alternative. Big data are a distinct "cultural, technological, and scholarly phenomenon" (2) centered on the application of machine learning algo-rithms to diverse, large-scale data. As clinics and hos-pitals generate huge amounts of electronic health record (EHR) data and systems like IBM's Watson system combine genomic data, published literature, and EHR data to guide cancer treatment (3), the pace, data sources, and methods for generating medical evidence are changing radically. Traditional clinical researchers rightly wonder whether, how, and why to engage with big data.

Data, Information, and Knowledge

Much of the excitement about big data methods is stoked by the explosive availability of diverse data that are "big" in volume, velocity, and variety. The EHR data of a typical middle-aged person are approximately the size of the collected works of Shake-speare. Mobile sensors can collect such varied data as heart rate, geolocation, and blood glucose levels with contact lenses (4). Social network and media data are a new window into patients' social contexts. However, data does not equal knowledge: Understanding the distinction among data, information, and knowledge is necessary to bridge the gap between big data and evidence-based medicine.

Data are raw observations that have limited value by themselves. What gives a raw observation value (for example, a hemoglobin A_{1c} level of 8.2%) is placing it in an interpretive context to yield information (for example, hemoglobin A_{1c} level of 8.2% is above the normal range). Data and information pertain to specific situations (such as a person, clinic, or country), whereas knowledge comprises general statements about the world that are useful for explaining, predicting, or guiding future action (for example, patients with high hemoglobin A_{1c} levels have diabetes and are at higher risk for cardiovascular illness). Knowledge may be explicit (such as statements in textbooks or guidelines) or tacit (such as diagnostic strategies of a master clinician) and is produced by the application of analytic methods on data and information. Therefore, statements of knowledge are claims for which evidence must be provided in the form of supportive data and analysis.

Evidence Is the Basis for a Claim of Knowledge

Evidence-based medicine and big data represent 2 very different approaches to producing evidence. In traditional EBM, a hypothesis is posed and data are collected through a study and then analyzed by using frequentist biostatistics to support a finding or a claim of knowledge. The classic randomized, controlled trial produces evidence on questions of causal inference (such as drug efficacy), whereas other study designs address diagnosis and prediction (such as diagnostic test studies or clinical prediction rules) or describe the natural history of disease (such as cohort studies). There is an amassed rich body of expertise on how to critically appraise claims of knowledge arising from these types of studies, and most researchers are well-versed in such concepts as selection bias and the benefits of randomization. The lingo and mindset associated with EBM are now deeply imbued in many generations of clinicians and clinical researchers.

Practitioners of big data share no such lingo with EBM. Data science practitioners come from a computational tradition that is driven by data rather than hypothesis testing. These methods work off raw observations and do not incorporate context knowledge into evidence production. Therefore, an algorithm may detect a pattern in a database but have no way of recognizing whether the result is true, spurious, or affected by bias. Therein lies the most critical difference between EBM and big data. Evidence-based medicine prioritizes the explicit control of biases in both data collection and analysis to maximize internal validity. In contrast, big data approaches rarely involve protocol-directed data collection but aim to maximize precision and external validity by the dictum of "more data are better than better data." The concept of bias, which requires the application of context knowledge to analysis, has no natural place in data-driven methods. Traditional researchers may find big data's epistemological approach heretical, but with the global market of big data analytics totaling $125 billion in 2015 (5), there is no walling off clinical research from these methods, nor should we want to because EBM and big data have complementary strengths.

Synergy Between 2 Ways of Knowing

The Appendix Figure shows how big data methods can map onto a taxonomy of study types familiar to most clinical researchers (6). For descriptive studies that describe a state of affairs, traditional survey and qualitative studies can be complemented by data mining large-scale data. For example, a description of attitudes toward human papillomavirus vaccination could

This article was published at www.annals.org on 26 January 2016.

be obtained through traditional survey methods on thousands of respondents (7) or could be gleaned through automated classification of positive and negative sentiment on 130 million English-language blog posts (8).

Big data methods offer expanded research power, especially for analytic studies aimed at classification, prediction, modeling, and simulation. Classification algorithms can act as diagnostic tests, classifying a patient as having or as not having a disease. For example, a classification algorithm trained on 2.1 million tweets was able to recognize Twitter users with depression with an accuracy of 70% and a positive predictive value of 74% (9). Predictive analytics similar to those used to predict whether a borrower will default on a loan can be applied to predicting disease outcomes. Modeling and simulation methods similar to those used for climate modeling can be applied to modeling cancer growth or a contagion. For causal inference, the randomized trial remains the gold-standard study design, but traditional, nonrandomized designs and causal learning algorithms may offer sufficient evidence in some situations.

Thus, EBM practitioners would do well to seek data science partners to exploit the availability of new, large-scale, diverse data and to enlarge their tool kit with machine learning methods that may offer less expensive, quicker, and more powerful approaches to generating evidence in some circumstances. Big data scientists, who often come from outside the health field, would do well to partner with clinical researchers who have the disease knowledge to adjust for sources of bias and to recognize spurious signals. Evidence-based medicine needs the computational power of big data, and big data needs the epistemological rigor of EBM. Combining these 2 ways of knowing offers the best path for enlarging and strengthening the knowledge base of clinical medicine.

From University of California, San Francisco, San Francisco, California.

Presented in part at the 3rd Annual Cochrane Lecture, Vienna, Austria, 4 October 2015 (available at www.youtube.com/watch?v=RgOgcs95fRk).

Disclosures: Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M15-2970.

Requests for Single Reprints: Ida Sim, MD, PhD, Division of General Internal Medicine, University of California, San Francisco, 1545 Divisadero Street, Suite 308, San Francisco, CA 94143-0320; e-mail, ida.sim@ucsf.edu.

Author contributions are available at www.annals.org.

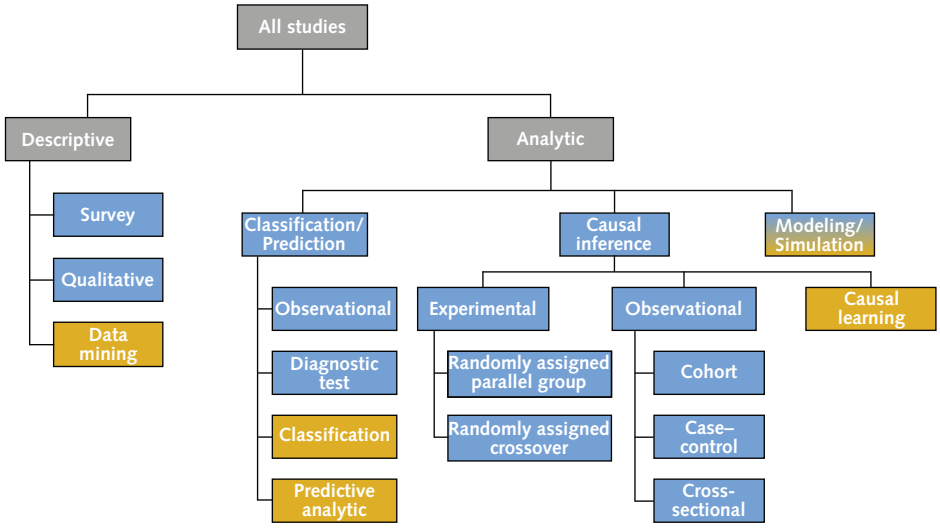
Ann Intern Med. doi:10.7326/M15-2970

References

1. **Evidence-Based Medicine Working Group.** Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA. 1992;268:2420-5. [PMID: 1404801]
2. **Boyd D, Crawford K.** Critical questions for big data. Information, Communication & Society. 2012;15:662-79.
3. IBM Watson for Oncology. Accessed at www.ibm.com/smarterplanet/us/en/ibmwatson/watson-oncology.html on 9 December 2015.
4. **Savov V.** Google signs deal to put sensors directly on your eye. The Verge. 15 July 2014. Accessed at www.theverge.com/2014/7/15/5900871/google-and-novartis-smart-contact-lens-partnership on 9 December 2015.
5. **Press G.** 6 Predictions for the $125 Billion Big Data Analytics Market in 2015. Forbes. 11 December 2014. Accessed at www.forbes.com/sites/gilpress/2014/12/11/6-predictions-for-the-125-billion-big-data-analytics-market-in-2015 on 9 December 2015.
6. **Centre for Evidence-Based Medicine.** Study Designs. Accessed at www.cebm.net/study-designs on 8 January 2016.
7. **Zhao FH, Tiggelaar SM, Hu SY, Zhao N, Hong Y, Niyazi M, et al.** A multi-center survey of HPV knowledge and attitudes toward HPV vaccination among women, government officials, and medical personnel in China. Asian Pac J Cancer Prev. 2012;13:2369-78. [PMID: 22901224]
8. **Corley CD, Mihalcea R, Mikler AR, Sanfilippo AP.** Chapter 18: Predicting individual affect of health interventions to reduce HPV prevalence. In: Arabnia HR, Tran QN, eds. Software Tools and Algorithms for Biological Systems. New York: Springer Science+Business Media; 2011:181. (Advances in Experimental Medicine and Biology, Vol. 696.)
9. **De Choudhury M, Gamon M, Counts S, Horvitz E.** Predicting Depression via Social Media. Proceedings of the Seventh International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media, Boston, MA, 8–10 July 2013. Palo Alto, CA: Association for the Advancement of Artificial Intelligence Pr; 2013.

Author Contributions: Conception and design: I. Sim.
Drafting of the article: I. Sim.
Critical revision of the article for important intellectual content: I. Sim.
Final approval of the article: I. Sim.
Administrative, technical, or logistic support: I. Sim.

Appendix Figure. Taxonomy of traditional and big data study types.



Clinical studies include descriptive studies, which aim to describe a state of affairs, and analytic studies, which aim to quantify a relationship. Blue boxes represent traditional clinical study designs. Orange boxes represent examples of big data methods. Both traditional and big data methods are applicable to modeling and simulation. Adapted from reference 6.