Published in final edited form as:

Circ Cardiovasc Qual Outcomes. 2016 November ; 9(6): 679–682. doi:10.1161/CIRCOUTCOMES.116.003097.

Can Big Data Fulfill Its Promise?

Peter W. Groeneveld, MD, MS[1,2,3,4] and **John S. Rumsfeld, MD, PhD**[5,6,7]

[1]Department of Veterans Affairs' Center for Health Equity Research and Promotion, Michael J. Crescenz Veterans Affairs Medical Center, Philadelphia, PA

[2]Division of General Internal Medicine, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

[3]Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA

[4]Center for Cardiovascular Outcomes, Quality, and Evaluative Research, University of Pennsylvania, Philadelphia, PA

[5]University of Colorado School of Medicine, Aurora, CO

[6]Veterans Affairs Eastern Colorado Health System, Denver, CO

[7]American College of Cardiology, Washington, DC

Journal Subject Terms

Quality and Outcomes; Machine Learning

Keywords

analysis; prediction statistics; statistical analysis; database; computer-based model

Big data analytics are widely touted as a key innovation to improve health care. While the term "big data" is variably defined, it generally implies the application of advanced statistical analyses, with names such as "machine learning," "artificial intelligence," or "cognitive computing," to data sources that greatly exceed the size and complexity of databases traditionally used for health care analyses.[1] Somewhat counterintuitively, data volume alone does not qualify an analysis as "big data." The term more appropriately refers to analytic and computational techniques, combined with information technology innovations, that were specifically developed to yield insights from the very large quantities of data that are increasingly common in the "digital economy." Big data analytics offer the promise of turning large amounts of data into superior predictive models that can be used to improve health care quality and outcomes.

Address for Correspondence: Peter W. Groeneveld, MD, MS, Perelman School of Medicine, University of Pennsylvania, 1201 Blockley Hall, 423 Service Drive, Philadelphia, PA 19104-4155, Tel: 215-898-2569, Fax: 215-573-8778, petergro@upenn.edu.

Conflict of Interest Disclosures: Dr. Rumsfeld is Chief Innovation Officer for the American College of Cardiology. The opinions expressed in this paper are the authors' own, and do not represent the official views of the U.S. Department of Veterans Affairs or the American College of Cardiology.

Since its inception, *Circulation: Cardiovascular Quality and Outcomes* has focused on using scientifically rigorous data analyses to improve health care,[2] thus the journal is a natural home for scientific investigations using big data analytic techniques. The current Special Issue highlights several high-quality studies that reveal the depth and breadth of these new methods and points toward a future where big data analytics are incorporated into the daily practice of all clinicians, much as big data analytics routinely impact the everyday online experiences of millions of users of Google, Amazon, and other internet-based companies.

While the term "big data" is a relatively new invention,[1] many of the techniques of big data analytics have been in existence for decades, and the field of biomedical informatics has been critical to their development.[3] Several recent phenomena have converged to move big data analytics to the forefront of health care, including the widespread adoption of electronic medical records and subsequent digitization of large volumes of health care data, advancements in computing and information technology capabilities, and a cadre of "data scientists" who are migrating to the health care sector and who bring know-how and experience gained in other data-intensive industries. However, it is not yet clear how big data should change health care delivery. Big data methods do not solve all data and analytic problems, and they also present unique challenges to the process of evaluating the methodological rigor and validity of data analysis and results. Health care interventions based on big data analytics will need thorough evaluation to prove their effectiveness, as new methods of analyzing data do not address the real-world challenges of implementing new health care practices and improving quality.

The Advantage of Big Data Analytics

The potential for big data analytics to transform health care is tantalizing, including the development of clinical prediction models that outperform the standard paradigm where a relatively small (e.g., <100) number of variables are pre-selected as potential predictors and an even smaller number of predictive variables (e.g., <10) are retained in a final prediction model. Cardiovascular medicine has been a leader in the clinical use of models developed using these techniques. Examples include the Goldman[4] and Detsky[5] criteria for preoperative cardiac risk assessment, the TIMI Risk Score for prognosis in acute coronary syndrome,[6] and the CHA_2DS_2-VASc score for stroke prediction in atrial fibrillation.[7] Conversely, big data analytics can potentially use thousands of variables, with tens of thousands of permutations, to produce dynamic predictive models that are continuously informed by newly collected information, as opposed to the static models generated by standard multivariable regression methods. Big data analytics are also adaptive to temporal and geographic variation in the data streams that are being analyzed, thus predictions are "tailored" to the time and place that they are generated.

The advantages of big data analytics are principally based on the following two premises:

- Premise #1. It is not possible to pre-specify all possible causal and/or associational pathways in a data environment with thousands (or hundreds of thousands) of variables. In fact, efforts at pre-specification inherently reduce the predictive ability of the data by artificially constraining the "choice set" of

predictors to a very limited subset of possible inputs.[8] By instead allowing computational algorithms to test a very large number of potential associations, the data are empowered to "speak for itself," and patterns of association may be identified that could not have been predicted in advance, even by the most knowledgeable investigator(s).

- Premise #2. To maximize operational usefulness, clinical prediction models must adapt to changes in the data over both space and time.[9] Logistic regression models (e.g., CHA_2DS_2-VASc) produce the same prediction of atrial fibrillation stroke risk, given the same values on the model's seven individual input variables, whether a patient is a low-income resident of West Philadelphia receiving care in a public health clinic or a wealthy resident of Atherton, California receiving "concierge care" from her private physician. The model's prediction that these two hypothetical patients have identical stroke risk is almost certainly wrong, as stroke risk is undoubtedly determined by many more factors than the seven CHA_2DS_2-VASc inputs. In contrast, big data analytics that can "learn on the job" and adjust predictions based on variables unique to the time and place at which the data are generated may have a significant advantage in predictive accuracy. Furthermore, as risk factors and their relationship to adverse outcomes inevitably change over time, big data analytic models can 'naturally' adapt, whereas static prediction models derived from a single derivation dataset cannot.

The Blackjack Conundrum

While big data analytics have undeniable potential advantages, advocates frequently overlook some of the basic observational data analysis problems that are not solved by better algorithms, more computational power, and petabytes of data.[1] The game of blackjack provides a simple example of the two fundamental problems with any exercise in predictive analysis, regardless of dataset size or computational power or methods—namely, the problems of *unobservability* and *randomness*. Beating the "house" with 95% (or even 55%) frequency in blackjack would require both accurate knowledge of the identity of face-down cards, as well as knowledge of the cards that are to be dealt next from the deck. But since face-down cards are (obviously) unobservable, and the cards coming out of the deck (in the absence of card-counting techniques) are essentially random, no amount of data, analytic complexity, or machine learning can increase a good blackjack player's long-term win (i.e., predictive) percentage above approximately 40%.[10] The same two issues of unobservability and randomness likewise bedevil prediction in health care. Not infrequently, important predictors of health outcomes may be unobservable—defined as neither measured directly nor measured by some combination of proxy measures, analogous to the face-down card in blackjack. It is tempting to assume that in the comprehensive electronic medical record era, all important predictors of health outcomes are captured in the data, but this is simply not true. Second, the impact of random events in determining many health outcomes cannot necessarily be ignored, and in some instances is considerable. The occurrence of most adverse medical events is likely influenced in part by a multitude of essentially random processes that defy prediction, even with a super-abundance of data. Thus, we should not be

surprised when big data analytics fail to provide predictive information that greatly surpasses standard statistical methods in predictive accuracy.[3, 11]

Even in the age of the electronic health record, data quality remains a challenge to precise prediction at the point of care. Missing and/or inaccurate data, particularly on patient-centered outcomes, imperils the ability of any predictive method to accurately classify patients. While machines can be programmed to "learn" from their previous mistakes, it is important to recognize that big data cannot fix bad data.

Is "Data Science" Science?

In addition to fundamental data problems that remain unsolved by big data analytics, the complexity and non-static nature of these methods are a challenge for traditional scientific reporting in journals. The fundamental idea behind scientific publication is the axiom of reproducibility, i.e., a scientific team reports their methods and results in a manner that would permit another scientific team to reproduce the results of the study. By its nature, big data analytics defies this axiom by (1) the analytic process's continually adapting to changes in the data, thus there is no "model" to present in a scientific report, (2) the use of complex computational algorithms that are difficult to describe to the uninitiated in the constrained space of a typical scientific methods section, and (3) an ever-expanding lexicon of big data analytic techniques (ridge regression, the lasso, generalized additive models, random forests, boosting, support vector machines, k-means clustering, etc.) that are unfamiliar to a large fraction of most medical journals' readership, even to those that are familiar with the basic tools of statistical analysis. In these ways, "data science" is more akin to engineering—in which the focus is typically on solving a set of unique design problems—rather than to biomedical sciences where the focus is on understanding the workings of the natural world.

This difference poses a unique challenge to journal editors, reviewers (including statistical experts), and readers. How can the methodological quality of big data analytics be externally evaluated? Some of the most rigorous published scientific reports using big data analytics have provided lengthy appendices with careful documentation and voluminous code (R, Python, MATLAB, etc.), but given the exponentially expanding syntax and variety of programming languages and techniques, even the most savvy reviewers will not be "fluent" in all languages. Nor are editors, reviewers, and readers likely to have the time necessary to deconstruct thousands of lines of code and assess its appropriateness for the task. The alternative to this level of review is even less appealing—namely, analytic methods written as a terse stream of statistical jargon that is incomprehensible to most readers and that reveals little of the inner workings of the modeling process, i.e., the metaphorical "black box." To ensure the integrity of the scientific peer-review process of big data analytics, novel methods of evaluation will need to be developed.

Concerns about the transparency of big data analytics are not confined to journal editors. In an era where health care quality report cards are typically reported as ratios of observed to predicted adverse events, it is important for health care providers, policymakers, and payers to clearly understand the methods that produce the "predicted events" denominators. Big data methods may increase the accuracy of the predicted number of adverse events per

surgeon or hospital, but if the methods are buried in machine learning algorithms, this will not enhance stakeholder trust in the accuracy and fairness of quality ratings.

Beyond transparency, as big data analytics are incorporated into clinical practice it will be vital to compare the clinical outcomes resulting from their use versus outcomes from previously developed clinical decision support tools. While published comparisons of various predictive techniques should appropriately use the statistical methods (C-statistics, etc.) developed to quantify the performance of competing classification schemes, it is also important to recognize that numerical improvement in prediction performance may not necessarily translate into better health outcomes for patients. Thus, beyond demonstrating better C-statistics, evidence that big data predictive models actually improve health outcomes when applied in real-world clinical settings, while not producing unintended consequences that harm patients, is essential. The stakes in health care decision-making are arguably higher than the use of big data analytics applied in many other fields. The promise of big data analytics is not a valid excuse to bypass the requirement that innovations must be scientifically proven to be effective in practice. In this regard, two additional issues are worthwhile to highlight as big data analytics emerge in health care: causal inference and quality improvement. Both are critical to the operational, effective use of data to improve health outcomes, yet neither are necessarily enhanced by new analytic methods.

In Healthcare, Causal Inference Matters

As with any observational data analysis, associations between predictors and outcomes in big data analytic models cannot be determined to be causal or non-causal. The risk of identifying spurious associations (classic examples include sunspot activity correlated with the stock market[12] and coffee consumption associated with pancreatic cancer[13]) is higher when the number of associations being tested is greater, and when there is no prior knowledge of causal pathway plausibility guiding the selection of model predictors. This is important because in many (most) instances, clinicians, administrators, and policy-makers want to know not only *who* the patients in their clinics/hospitals/health systems are that will have an adverse clinical outcome, but also *why* the adverse outcome occurred—i.e., the identity of causal variables and whether those variables are modifiable. So while it is important to predict, for example, which intensive care unit patients are highly likely to develop bloodstream infections due to central intravenous catheters, it is also essential to identify key modifiable causal factors that could reduce the future incidence of bloodstream infections.

Quality Improvement Is More Than Analytics

The complexity of health care delivery creates barriers to the appropriate use of information. There is no assurance that the availability of new, accurate predictive information about patients will transform the process of health care so that adverse events are avoided and the probability of good outcomes is maximized. Big data analytics has the potential to generate useful predictive information that could be delivered to clinicians in a timely fashion (via electronic medical record pop-ups, text messages, email, etc.), but that information may be overlooked by clinicians bombarded with too much information, or ignored because the

appropriate action in response to the information is neither obvious nor simple.[14] Thus, better predictive information is necessary but not sufficient for the optimization of health care outcomes, and lessons from the relatively new field of implementation science will be essential components of leveraging big data analytics.

In summary, the rise of big data analytics in health care settings presents an exciting opportunity to leverage the power of increasingly voluminous health care data in ways that were simply impossible as recently as ten years ago. However, it is critical to recognize that the fundamental pitfalls of observational data analysis cannot be ignored, and in fact the risks of such pitfalls demand rigorous scientific testing and novel methods for peer review of big data analytic models. Neither enthusiasm for the potential of big data analytics to transform care, nor the complexity of big data methods, should obviate the need for rigorous scientific evaluation. Big data analytics are fundamentally a tool that will require, like any other medical intervention, clinical integration with quality improvement activities to have a meaningful, positive impact on health and health care. The field of cardiovascular outcomes research should embrace these new techniques, rigorously assess the scientific studies using these methods, and guide policymakers, clinicians, and patients on their meaningful implementation.

References

1. Mayer-Schönberger, V., Cukier, K. Big Data: A Revolution that will Transform how we Live, Work, and Think. Boston: Houghton Mifflin Harcourt; 2013.
2. Krumholz HM. Outcomes research: myths and realities. Circ Cardiovasc Qual Outcomes. 2009; 2:1–3. [PubMed: 20031804]
3. Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. Ann Thorac Surg. 1997; 63:1635–1643. [PubMed: 9205161]
4. Goldman L, Caldera DL, Nussbaum SR, Southwick FS, Krogstad D, Murray B, Burke DS, O'Malley TA, Goroll AH, Caplan CH, Nolan J, Carabello B, Slater EE. Multifactorial index of cardiac risk in noncardiac surgical procedures. N Engl J Med. 1977; 297:845–850. [PubMed: 904659]
5. Detsky AS, Abrams HB, Forbath N, Scott JG, Hilliard JR. Cardiac assessment for patients undergoing noncardiac surgery. A multifactorial clinical risk index. Arch Intern Med. 1986; 146:2131–2134. [PubMed: 3778043]
6. Antman EM, Cohen M, Bernink PJ, McCabe CH, Horacek T, Papuchis G, Mautner B, Corbalan R, Radley D, Braunwald E. The TIMI risk score for unstable angina/non-ST elevation MI: A method for prognostication and therapeutic decision making. JAMA. 2000; 284:835–842. [PubMed: 10938172]
7. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010; 137:263–272. [PubMed: 19762550]
8. Anderson, C. The end of theory: the data deluge makes the scientific method obsolete. Available at http://www.wired.com/2008/06/pb-theory/. Accessed June 1, 2016
9. Kanjilal, PP., Institution of Electrical Engineers. Adaptive prediction and predictive control. Stevenage, Herts., U.K.: P. Peregrinus on behalf of Institution of Electrical Engineers; 1995.
10. Baldwin R, R CWE, Maisel H, McDermott JP. The optimum strategy in blackjack. J Am Stat Assoc. 1956; 51:429–439.
11. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. Nat Rev Cardiol. 2016; 13:350–359. [PubMed: 27009423]

12. Collins CJ. An inquiry into the effect of sunspot activity on the stock market. Finance Analyst J. 1965; 21:45–56.
13. MacMahon B, Yen S, Trichopoulos D, Warren K, Nardi G. Coffee and cancer of the pancreas. N Engl J Med. 1981; 304:630–633. [PubMed: 7453739]
14. Wachter, RM. The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age. McGraw-Hill Education; New York: 2015.