The Inevitable Application of Big Data to Health Care

Travis B. Murdoch, MD, MSc
Allan S. Detsky, MD, PhD

THE AMOUNT OF DATA BEING DIGITALLY COLLECTED AND stored is vast and expanding rapidly. As a result, the science of data management and analysis is also advancing to enable organizations to convert this vast resource into information and knowledge that helps them achieve their objectives. Computer scientists have invented the term *big data* to describe this evolving technology. Big data has been successfully used in astronomy (eg, the Sloan Digital Sky Survey of telescopic information), retail sales (eg, Walmart's expansive number of transactions), search engines (eg, Google's customization of individual searches based on previous web data), and politics (eg, a campaign's focus of political advertisements on people most likely to support their candidate based on web searches).

In this Viewpoint, we discuss the application of big data to health care, using an economic framework to highlight the opportunities it will offer and the roadblocks to implementation. We suggest that leveraging the collection of patient and practitioner data could be an important way to improve quality and efficiency of health care delivery.

Widespread uptake of electronic health records (EHRs) has generated massive data sets. A survey by the American Hospital Association showed that adoption of EHRs has doubled from 2009 to 2011,[1] partly a result of funding provided by the Health Information Technology for Economic and Clinical Health Act of 2009. Most EHRs now contain quantitative data (eg, laboratory values), qualitative data (eg, text-based documents and demographics), and transactional data (eg, a record of medication delivery). However, much of this rich data set is currently perceived as a by-product of health care delivery, rather than a central asset to improve its efficiency.

The transition of data from refuse to riches has been key in the big data revolution of other industries. Advances in analytic techniques in the computer sciences, especially in machine learning, have been a major catalyst for dealing with these large information sets. These analytic techniques are in contrast to traditional statistical methods (derived from the social and physical sciences), which are largely not useful for analysis of unstructured data such as text-based documents that do not fit into relational tables. One estimate suggests that 80% of business-related data exist in an unstructured format.[2] The same could probably be said for health care data, a large proportion of which is text-based.

In contrast to most consumer service industries, medicine adopted a practice of generating evidence from experimental (randomized trials) and quasi-experimental studies to inform patients and clinicians. The evidence-based movement is founded on the belief that scientific inquiry is superior to expert opinion and testimonials. In this way, medicine was ahead of many other industries in terms of recognizing the value of data and information guiding rational decision making. However, health care has lagged in uptake of newer techniques to leverage the rich information contained in EHRs.[3] There are 4 ways big data may advance the economic mission of health care delivery by improving quality and efficiency.

First, big data may greatly expand the capacity to generate new knowledge. The cost of answering many clinical questions prospectively, and even retrospectively, by collecting structured data is prohibitive. Analyzing the unstructured data contained within EHRs using computational techniques (eg, natural language processing to extract medical concepts from free-text documents) permits finer data acquisition in an automated fashion. For instance, automated identification within EHRs using natural language processing was superior in detecting postoperative complications compared with patient safety indicators based on discharge coding.[4] Big data offers the potential to create an observational evidence base for clinical questions that would otherwise not be possible and may be especially helpful with issues of generalizability. The latter issue limits the application of conclusions derived from randomized trials performed on a narrow spectrum of participants to patients who exhibit very different characteristics.

Second, big data may help with knowledge dissemination. Most physicians struggle to stay current with the latest evidence guiding clinical practice. The digitization of medical literature has greatly improved access; however, the sheer

Author Affiliations: Division of Gastroenterology, University of Calgary, Calgary, Alberta, Canada (Dr Murdoch); Institute for Health Policy, Management and Evaluation, and Department of Medicine, University of Toronto, and Department of Medicine, Mount Sinai Hospital and University Health Network, Toronto, Ontario, Canada (Dr Detsky).
Corresponding Author: Allan S. Detsky, MD, PhD, Department of Medicine, Mount Sinai Hospital, 600 University Ave, Ste 429, Toronto, ON M5G 1X5, Canada (adetsky@mtsinai.on.ca).

number of studies makes knowledge translation difficult. Even if a clinician had access to all the relevant evidence and guidelines, sorting through that information to develop a reasonable treatment approach for patients with multiple chronic illnesses is exceedingly complex. This problem could be addressed by analyzing existing EHRs to produce a dashboard that guides clinical decisions. This approach is being used in the collaboration between IBM's Watson supercomputer and Memorial Sloan-Kettering Cancer Center to help diagnose and propose treatment options for patients with cancer. The big data approach differs from traditional decision support tools in that suggestions are drawn from real-time patient data analysis, rather than solely using rule-based decision trees. For example, longitudinal diagnostic data have been shown to predict a patient's risk of a future diagnosis of domestic abuse.[5] Data-driven clinical decision support tools could also lead to cost savings and help with appropriate standardization of care. Just as purchasers receive messages from Amazon (eg, customers like you also bought this book), clinicians may receive messages that inform them about the diagnostic and therapeutic choices made by respected colleagues facing similar patient profiles.

Third, big data may help translate personalized medicine initiatives into clinical practice by offering the opportunity to use analytical capabilities that can integrate systems biology (eg, genomics) with EHR data.[6] The Electronic Medical Records and Genomics Network does so by using natural language processing to phenotype patients, in an effort to streamline genomics research.

Fourth, big data may allow for a transformation of health care by delivering information directly to patients, empowering them to play a more active role. The current model stores patients' records with health care professionals, putting the patient in a passive position. In the future, medical records may reside with patients. Big data offers the chance to improve the medical record by linking traditional health-related data (eg, medication list and family history) to other personal data found on other sites (eg, income, education, neighborhood, military service,[7] diet habits, exercise regimens, and forms of entertainment), all of which can be accessed without having to interview the patient with an exhaustive list of questions. By doing so, big data offers a chance to integrate the traditional medical model with the social determinants of health in a patient-directed fashion. Public health initiatives to reduce smoking and obesity could perhaps be delivered more efficiently in this way by targeting their messages to the most appropriate people based on their social media profiles.

There are several important barriers to the widespread adoption of big data in health care.[3] At first glance there seem to be no strong incentives or champions for its use within clinician groups or hospitals. However, the same could be said for the patient safety movement 10 years ago, which has since been widely embraced. The era of value-based payments may provide the needed stimulus. Moreover, there are many champions for big data within the parts of the health sector that do not deliver direct individual patient care such as health service researchers, pharmaceutical companies, public health and other government organizations.[3]

Moreover, there will be considerable privacy concerns, which will require solutions similar to, and perhaps even more extensive than, those required to protect confidential financial data in other sectors. Also, current EHR platforms are fragmented and have limited interoperability.[3] However, the same issue arises in other sectors of the economy where companies use multiple systems for recording data. In addition, just as EHRs and electronic order entry have been shown to produce considerable safety issues (eg, inadvertant selection of incorrect medications), big data will likely do so as well. Overreliance on electronic systems is a problem for all sectors. For example, automated piloting systems for airplanes sometimes require human override, necessitating the need for simulation training in case of system failure.

Economic theory describes the quantitative conversion of 3 kinds of inputs (capital, labor, and raw materials) into outputs (goods and services). This technical relationship is known as a production function. As technology progresses this quantitative relationship changes, usually requiring fewer inputs to yield the same or more output. The current revolution in data management makes it clear that a fourth kind of input, information, will become just as important as these other inputs in the future of many industries. As such, the application of big data to health care is inevitable. The first information technology revolution in medicine was the digitization of the medical record. The second is surely to leverage the information contained therein and combine it with other sources. Big data has the potential to transform medical practice by using information generated every day to improve the quality and efficiency of care.

Conflict of Interest Disclosures: Both authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.
Additional Contributions: We thank Robert M. Wachter, MD, University of California, San Francisco, and Robert H. Brook, MD, ScD, University of California, Los Angeles, for their comments on an earlier draft. These persons received no compensation for their contributions.

REFERENCES

1. Charles D, Furukawa M, Hufstader M. *Electronic Health Record Systems and Intent to Attest to Meaningful Use Among Non-federal Acute Care Hospitals in the United States: 2008-2011*. Office of the National Coordinator for Health Information Technology; 2012. http://www.healthit.gov/media/pdf/ONC_Data_Brief_AHA_2011.pdf. Accessed December 3, 2012.
2. Grimes S. Unstructured data and the 80 percent rule. Clarabridge BridgePoints, 2008. http://clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551. Accessed December 3, 2012.
3. Safran C, Bloomrosen M, Hammond WE, et al; Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *J Am Med Inform Assoc*. 2007;14(1):1-9.
4. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*. 2011;306(8):848-855.
5. Reis BY, Kohane IS, Mandl KD. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. *BMJ*. 2009;339:b3677.
6. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012;13(6):395-405.
7. Brown JL. The unasked question. *JAMA*. 2012;308(18):1869-1870.