Healthcare's "Big Data" Challenge

Julia Adler-Milstein, PhD; and Ashish K. Jha, MD, MPH

There is enormous enthusiasm for how "big data" can address persistent cost and quality deficiencies in the healthcare system. The notion is simple—analyzing the massive amount of clinical information that is newly available in digital format should enable groundbreaking insights: earlier detection of effective treatments, better targeted clinical decision support, real-time biosurveillance, and accurate predictions of who is likely to get sick. Indeed, it will take harnessing these data to create what the Institute of Medicine calls a "learning health system" in which we continuously identify and adopt new approaches to deliver better care at lower costs. Big data has generated even more excitement since the most recent election, when analysts were able to use vast amounts of disparate data from polls, economic indicators, and historical trends to predict election outcomes with astounding accuracy. Yet, the excitement about big data, and the analytics that it requires, appears to have gotten ahead of the reality. Without more specific attention to the challenges, an important tool for transforming healthcare will fail to deliver on its promise.

The first set of challenges is those that confront all big data efforts. We need continual technical advances to store and efficiently access the rapidly expanding amount of data. There are also challenges that are particularly salient in healthcare. Concerns about privacy and security are paramount, although these are increasingly being addressed by new authentication approaches and policies that better safeguard patient-identifiable data. The challenge that demands significantly more attention is ensuring that the data are not only big but that they are good.

Healthcare does not have a problem with big. The rate of electronic health record (EHR) adoption continues to climb in both inpatient and ambulatory settings. In 2011, EHRs were used to capture the clinical care in 13 million hospitalizations, 450 million outpatient visits, and countless pharmacies, laboratories, and other sites. And increasingly, patients are using devices to track their health and health-related behaviors which generate substantial data. We have achieved big with exponential growth ahead.

In this article
Take-Away Points / p538
www.ajmc.com
Full text and PDF

The problem in healthcare lies with the quality of the data. To derive insights from data, it is critical that they be accurate and relatively complete. When data are systematically biased through either errors or omissions, the correlations that give rise to new insights will be missed or spurious, and could result in misguided confidence or scarce resources dedicated to chasing down dead ends.

Healthcare should be disproportionately concerned about data quality for 2 reasons. First, a large fraction of data is entered by humans, who both intentionally and unintentionally introduce systematic errors. In most domains, data are a natural byproduct of our increasing reliance on technology. Internet searches, online purchases, and cell phone calls each create a treasure trove of data that can be mined for patterns (eg, flu symptom searches) or used for experimentation (do you buy more when the "cart" is in the upper right or the upper left?). The opportunities for data error are relatively limited and easily identified by algorithms.

In healthcare, critical clinical data—symptoms, physical signs, orders, and progress notes—still rely heavily on human entry and will do so for the foreseeable future. The opportunities to introduce errors are rife (eg, in structured fields, it is easy to select the option above or below the one that was intended).[1] Beyond simple data entry errors, there are bigger, systemic problems with clinical data in electronic format. Current reimbursement policies require extensive documentation and clinicians often respond by using templates that automatically populate large quantities of data or by using copy-paste features that propagate mistakes or outdated information. Further, physician documentation styles vary substantially, making errors and omissions difficult to identify. For instance, if a patient has an empty medication list, is the patient not taking any medications, has the patient not informed her physician about a medication she is taking, or has the physician chosen to document medications elsewhere?

Systematic data inaccuracies have no quick fixes. The most helpful, long-term solution would be for payers, especially Medicare, to relax documentation requirements so that providers can focus data entry on clinical, not billing, needs. Alternatively, there are promising technical solutions such as machine learning, a form of artificial intelligence that trains systems to make predictions about certain characteris-

Take-Away Points

- There is currently an explosion of electronic clinical data, which can be analyzed to glean new insights into how to improve overall health and healthcare.
- Two data quality issues—systematic data inaccuracies and data fragmentation—need to be addressed in order to ensure that insights from electronic clinical data are valid.
- These issues require specific attention from policy makers and practitioners, and may be lessened by promoting greater interoperability and reducing burdensome documentation requirements.

tics of data. While machine learning has proved successful for identifying missing diagnoses,[2] it is of limited use for critical clinical data like symptoms and physical exam findings. In addition, the richest source of clinical data, the clinicians' notes, remains largely beyond the reach of big data. There have been substantial improvements in natural language processing (NLP) to identify key information from clinical notes.[3] However, until the quality of those notes improve, it will be hard for NLP programs to glean important information. Continued investment in technical solutions will undoubtedly improve data accuracy, but without fundamental changes to how care is documented, we should be circumspect about our ability to rid data of systematic errors.

Even if we achieve perfect data accuracy, a second daunting challenge remains: data fragmentation. Incomplete data are common in clinical practice and reflect our highly fragmented healthcare system where patients see multiple clinicians whose EHRs do not communicate. Despite significant policy interest, we have yet to achieve any meaningful level of interoperability and without it, creating a comprehensive picture of patients' care will be nearly impossible. Incomplete data, like inaccurate data, can also lead to missed or spurious associations that can be wasteful or even harmful to patient care.

The solutions to address fragmented data are no easier than those to address inaccurate data. Achieving greater interoperability between electronic clinical systems has been pursued by policy makers for nearly 2 decades with little success.[4] While policy makers have recently renewed their efforts, their primary focus, moving specific pieces of clinical information (such as a laboratory test result) between individual healthcare providers, will do little to ensure that provider organizations have a comprehensive picture of the patient's care across all care sites. Many patients and privacy advocates are understandably concerned about efforts to aggregate data. However, with adequate de-identification and security safeguards, the risks of aggregation can be minimized and the benefits of better care at lower costs are substantial.

The potential for big data in healthcare is enormous and exciting. It is hard to find a delivery system that is not thinking about how to leverage EHR data, and researchers are eager to answer new types of questions. Realizing the most from our large national investment in health IT demands that we learn from the newly available data. Doing so requires that we understand the issues of data quality and address them effectively. The solutions are not easy. However, ignoring these challenges could quickly lead us from the hope for big data to the disappointing and wasteful results of bad data.

Author Affiliations: From School of Information (JA-M), University of Michigan, Ann Arbor, MI; Harvard School of Public Health (AKJ), Boston, MA.

Funding Source: None.

Author Disclosures: The authors (JA-M, AKJ) report no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article.

Authorship Information: Concept and design (JA-M); drafting of the manuscript (JA-M); critical revision of the manuscript for important intellectual content (JA-M, AKJ); administrative, technical, or logistic support (AKJ); and supervision (AKJ).

Address correspondence to: Julia Adler-Milstein, PhD, University of Michigan, School of Information, 4376 North Quad, Ann Arbor, MI 48109. E-mail: juliaam@umich.edu.

REFERENCES

1. Koppel R, Metlay JP, Cohen A, et al. Role of computerized physician order entry systems in facilitating medication errors. *JAMA*. 2005; 293(10):1197-1203.

2. Erraguntla M, Gopal B, Ramachandran S, Mayer R. Inference of missing *ICD-9* Codes using text mining and nearest neighbor techniques. Paper presented at: System Science (HICSS), 2012 45th Hawaii International Conference; Maui, HI; January 4-7, 2012.

3. Jha AK. The promise of electronic records: around the corner or down the road? *JAMA*. 2011;306(8):880-881.

4. Adler-Milstein J, Jha AK. Sharing clinical data electronically: a critical challenge for fixing the health care system. *JAMA*. 2012;307(16): 1695-1696. ■