Big Data Applications in Clinical Medicine

Sunil Nair, MD, and Leo Anthony Celi, MD

Sunil Nair is a hospitalist in the Department of Medicine at Beth Israel Deaconess Medical Center in Boston.
ssnair@bidmc.harvard.edu

Leo Anthony Celi is clinical research director at the Laboratory of Computational Physiology, MIT Institute for Medical Engineering and Science, and an intensivist at Beth Israel Deaconess Medical Center.
leoanthonyceli@yahoo.com

ABSTRACT

The widespread adoption of electronic medical records has created new opportunities for clinical investigation using big data techniques. The potential for nuanced investigation across a full range of clinical questions is tremendous, contingent on the investment hospitals and health systems can make in big data infrastructure. Secondary analysis of electronic health records will enable the use of real patient data to assist clinical decision making, with the goal of eventually providing near-real time support for bedside encounters.

Clinicians and patients will derive value from data-driven decision making, while hospitals and health systems may see returns in quality, patient safety, and satisfaction. For big data analytics to achieve its potential in clinical medicine, issues of data structure, analytics staffing, funding, and data security will have to be addressed, but the future is bright and fertile for the application of big data to medical care.

WHY BIG DATA IN CLINICAL MEDICINE?

More than any other healthcare relationship, the patient-doctor interaction is considered paramount among the U.S. medical community. Today, we are in an era of evidence-based medicine, which favors using expert guidelines based on scientific research to create a “standard of care”; health system reform that combines physicians into larger and larger groups, eliminating individuality; and population health management, where clinicians are evaluated—and paid—based on the outcomes of their whole patient panel rather than individual success stories. Yet many physicians still consider the "bedside" encounter, on which most of clinical decision making is based, the crux of healthcare delivery. Even in the 21st century, physicians justify many medical decisions with the simple words, "my clinical impression of the patient was…".

In this environment, one can understand physicians' resistance to big data analytics. Few things are more foreign to us than to input a complex set of clinical data from various sources into a black-box algorithm in order to draw conclusions that could affect patient care. Even so, as has already occurred in nearly every other American industry—as well as parts of the healthcare industry, such as drug safety or value-based purchasing—big data analytics is coming to medical care.

This is not to say big data is a completely novel idea to physicians. For years, large clinical trials—some with thousands of patients—have yielded the clinical insights physicians use to make individual care decisions for everything from blood pressure management in the clinic to ventilator settings in the intensive care unit. Perhaps the novelty for physicians today is that their own clinical data and impressions—in the form of electronic health records (EHRs)—are being used to create the data sets that will answer clinical queries. The next iteration of big data analytics won't be the traditional a priori creation of databases to answer targeted research questions but the use of real patient data, created by millions of patient interactions in hospitals and health clinics, to answer real-time questions.

This is the goal of data scientists and clinicians around the country and increasingly around the globe—to derive meaningful insights from information collected in day-to-day clinical settings. Why not? The data already exists and continues to expand exponentially in the servers and cloud spaces that host electronic health records systems. These systems capture all manner of clinical information previously sequestered on paper charts, bedside monitors, and laboratory and hospital information systems. Demographic information, physiological signal recordings, imaging, even narratives from clinicians are available for exploration and analysis, if only researchers know where—and how—to look.

BIG DATA FINDINGS MAY RIVAL OLDER RESEARCH METHODS

The current gold standard for clinical investigation is the randomized clinical trial (RCT). RCTs are meticulously constructed to minimize all manner of bias and confounding—the distortion of association between an exposure and outcome by an extraneous variable—either through trial design or subsequent statistical analysis. The essence of an RCT is a direct comparison in outcomes between an "exposure" group—which receives a certain test or therapy—and a "control" group, which does not. Much rides on the results of these trials, which shape technology adoption, standards of care, and revenue streams sometimes worth millions (or billions) of dollars. Consequently, much consideration is given to eliminating any unwanted discrepancy between the groups that could muddle the analysis.

Unfortunately, such thoroughness can have trade-offs. Making the study population as "pure" as possible is an easy way to reduce unwanted variation. However, perhaps not surprisingly, RCTs have been criticized for not recruiting the elderly, women, children, and minority groups. RCTs often have very specific inclusion and exclusion criteria for their participants, but real-world patients have no such limitations on the number of concurrent illnesses they may suffer from or the healthcare settings where they may seek medical attention. RCTs also require a minimum size to achieve statistically significant results. This may preclude conducting studies on rare diseases or rare interventions when there are too few patients to achieve adequate statistical power and thereby draw meaningful conclusions from the study.

There are also circumstances under which an RCT may be unethical—such as when patients cannot give consent to participate or when placement into the control group could result in withholding lifesaving treatment in dire circumstances. Both elements are often true in the critically ill population, which we study in particular. As they are often unable to speak for themselves, these patients pose a moral quandary for the clinical researcher. In addition, the patient's clinicians can't often be "blinded"—intentionally made unaware whether the patient is getting a treatment so as not to bias study observations and findings.

For all these reasons, investigators are turning to other means of performing clinical research, namely the secondary analysis of electronic health records data. Unlike in years past, they are not reviewing one or a few records but thousands—even tens of thousands—for useful conclusions. Such an analysis is considered secondary because the purpose of electronic medical records has often been not to facilitate clinical research but rather to meet administrative and documentation requirements (one challenge, among many, to be discussed below). Yet the clinical data exists, and forward-looking researchers and organizations are harnessing it to advance medical science.

In fact, the findings of large observational trials (which are essentially similar to big data analysis of electronic health records) and RCTs may be comparable. Work by John Ioannidis, MD, of Stanford University, and colleagues (2001) found significant correlation between the results of observational trials and RCTs across many topics in medicine, a finding repeated by a larger Cochrane Database study in 2014 (Anglemyer, Horvath, and Bero, 2014). Though there are no direct comparisons of the costs of RCTs versus secondary analysis of electronic health records, the latter piggybacks on existing infrastructure, while RCTs incur costs from patient recruitment, clinician and investigator compensation, and the publication and dissemination of discoveries.

NEW APPROACHES YIELD NOVEL INSIGHTS

One example of successful big data analytics in medicine is the study of Vioxx by Graham and colleagues (2005), one of the first to find that patients who took rofecoxib (the generic name for Vioxx) were at increased risk of serious complications compared to those who took other nonsteroidal anti-inflammatory drugs (NSAIDs). This finding was powered by 2,302,029 person-years of data drawn from the medical record system of Kaiser Permanente of California, an integrated health system that treats millions of patients yearly.

Though the researchers likely did not set out to prove the power of big data analytics as their primary objective, their findings are a powerful vindication of the use of electronic health records for clinical research and patient safety. An RCT followed and verified the association of Vioxx with serious side effects that led to death. Without the analysis of a large electronic health records database, it is unlikely Vioxx would have had the same level of scrutiny as it subsequently received.

The real prize is to generate actionable information that could alter a patient's immediate hospital course.

Pharmaceutical and medical device companies rarely create databases for post-market surveillance, which may reveal adverse reactions and interactions not discovered in the Phase III trials (where the efficacy of drugs and other medical interventions in humans is evaluated). It is left for researchers to analyze clinical information systems to verify claims of drug safety (or reports of side effects); the larger the information system, the greater the researchers' ability to detect unintended side effects across a more diverse population, many of whom may have been excluded during initial studies.

Merck, the maker of Vioxx, set aside billions to pay claims to thousands of patients who had taken its medication, and ultimately settled with state and federal health authorities for a misdemeanor charge related to its marketing. This is an example of big data analytics benefitting a whole population: a potentially harmful medication was withdrawn from the market. Although it is not clear whether individual physicians voluntarily reduced their prescribing of Vioxx prior to its recall as more damaging information on the drug was published, there was a great deal of physician commentary about the consequences of direct-to-consumer (DTC) advertising and transparency in medical publishing in the wake of Vioxx.

This is a big data analytics success story in many ways, but not quite the holy grail for clinical investigators and their big data scientist associates. The real prize for them is to bring big data analytics to bear on individual patient cases, in real time, to generate actionable information that could alter a patient's immediate hospital course. At least weekly (if not daily) in a tertiary care center, a question arises with no immediately apparent answer in current medical literature, as extensive as that literature has become. The next best proxy is to see how prior patients with similar demographics, disease states, and diagnostic and therapeutic options have done, which is to say, performing a massive data search, eliminating noise and addressing confounding by indication (adjusting the analysis for patient or disease characteristics), and arriving at meaningful conclusions—big data analytics, in short.

Recent publications suggest we may be on the cusp of achieving such capability, at least in select centers. When confronted with the decision whether to thin the blood of a child with a complicated illness to prevent blood clots, physicians at Stanford found no studies pertinent to their case, nor sufficient consensus among their colleagues to inform their decision. They were fortunate, however, to have access to Stanford's

STRIDE platform—a text search capability overlaid on Stanford's patient data warehouse. Within hours, they were able to find a cohort of patients roughly matching the one under their care and identify this cohort's risk of developing blood clots (which turned out to be quite high!). Based on this analysis, the decision was made to administer blood thinners within hours of the patient's admission, given the presumed high risk for clotting, (Frankovich, Longhurst, and Sutherland, 2011)

> The biopharmaceutical industry alone invested $10 billion in clinical trials–or approximately $9,090.91 per patient–in 2013.

Did this save the patient's life? It's impossible to say, the authors admit. No blood clots developed, but there's no telling if they would have occurred in the absence of blood thinners. However, the point of this case is how big data analytics empowered bedside physicians to provide care for a rare patient condition almost on the fly. Physicians at tertiary-level medical centers—to whom some of the most complex medical cases are referred—regularly confront such conundrums but often without the assistance of being able to query past cases.

Instead, in the absence of substantive medical studies (all of which are population-level data anyway and not necessarily reflective of the individual patient being cared for), physicians are often forced to rely on personal impressions, those of consultants who happen to be on-service that week, or, if they are lucky, chance encounters with colleagues or field experts from which a consensus may be derived. Big data analytics is potentially a big step up in supporting medical decision making for actual patient care.

WHO PAYS FOR THIS, AND OTHER CHALLENGES

Mentioned above are some of the limitations of RCTs, including their expense. Although RCTs will vary in cost by study design, diagnostic or therapeutic tool being investigated, number of study sites, and so on, every RCT will have startup, recurring, and closure expenditures, much like any entrepreneurial venture. Even with existing infrastructure, data definition, and data storage, tasks such as patient recruitment, clinician engagement, diagnostic and therapeutic administration, data analysis, and publication have to be done essentially from scratch for each RCT. (In fairness, RCT data can later become the substance of subgroup analyses or meta-analyses, saving costs while generating new conclusions.) In 2013, the biopharmaceutical industry alone invested $10 billion in clinical trials, with a total reported 1.1 million patients enrolled—or approximately $9,090.91 spent per patient (Battelle Technology Partnership Practice, 2015).

Observational studies performed via secondary analysis of electronic health records data could disrupt this dynamic, at least for certain applications and in particular care settings. That is not to say that big data analytics on EHR data is free and easy—in fact, the funding stream for such analysis may be less secure and the data potentially more difficult to work with. Consider that EHRs are implemented by hospitals and health systems for care organization, documentation, and billing requirements. Data is structured to fulfill these requirements—that *x* organ systems have been reviewed, *y* problems have been assessed, or *z* billing codes have been input. These information systems were not designed to inform clinical decision making by facilitating analysis of aggregated patient data.

Compounding the issue is who will be pulling the data and analyzing it. Those employed by the hospital to maintain the EHRs are not paid to participate in medical research and may not understand the research terminology or methodology of clinicians. Frankly, the skills of an IT/clinical data technician are not the same as a data scientist. After the relevant data is retrieved, someone with specialized training and skills has to actually preprocess and perform exploratory analyses to ascertain the accuracy of that data. These people (and any additional staff) would need to be paid, but this doesn't fit into traditional hospital budgets for clinical or administrative operations.

We are fortunate at the Beth Israel Deaconess Medical Center (BIDMC) to be part of the Medical Information Mart in Intensive Care (MIMIC)—a collaboration with the Laboratory of Computational Physiology (LCP) at the Massachusetts Institute of Technology, which is funded by the National Institute for Biomedical Imaging and Bioengineering. Using BIDMC intensive care unit data, physicians and data scientists have created an immense research resource of de-identified clinical information (where identifiable data elements such as names, addresses, and phone numbers have been removed).

In fact, the database is open access and free to use, provided one has completed an online training course on the rules and ethics of working with human subjects' data. The core staff that maintain the database is supported by federal research funding, an advantage that not all hospitals or health systems have. In fact, despite having their own electronic records systems, researchers from other institutes continue to use MIMIC precisely because the internal barriers to acquiring similar data at their own centers—such as funding and staff time—are too high.

Because MIMIC is an open-access database for those interested in critical care research, data entered into it has to be processed to protect patient privacy. MIMIC exists separately from the parent BIDMC clinical systems. Apart from removing the identifiable information mentioned above, dates are shifted randomly (but consistently within each patient to preserve durations of treatments and lengths of stay). Other efforts to create databases for big data analytics using patient data will similarly have to bear in mind issues of data security and patient protection. De-identification and controlling access are potential solutions but, again, ones that require resources to implement.

THE PROMISE OF BIG DATA IN CLINICAL MEDICINE

Big data analytics of clinical information is a promising frontier for medical research as well as clinical decision support for physicians and other allied health providers. Just using the MIMIC database, members of LCP and collaborators from around the world are looking at questions ranging from comparative effectiveness of drugs and other interventions to operations research and development of probabilistic models of critical illness. This is just using the one database.

Of course, this approach is not without limitations and challenges. It is unlikely, for instance, that professional societies and regulatory organizations would accept anything less than an RCT for safety and efficacy data to approve a new drug, device, or other diagnostic or therapeutic tool. However, there are times when RCTs may not be possible or would be too costly to perform. Similarly, and on a smaller scale, linking clinical questions with big data resources and tools may assist physicians in areas where traditional medical evidence does not exist.

Just as physicians reach out today to pharmacists, nutritionists, and physical therapists to provide comprehensive patient care, one can imagine a future clinician reaching out to a data scientist to perform a query on a database constructed from historical clinical data—or even real-time clinical data if adequate protections can be put in place. Perhaps the physician is interested in predicting response to a specific treatment among a complicated subgroup of patients, or estimating the risk of harm from a certain intervention. The query is run and returned to the physician within minutes or hours and can then be used at the bedside or to request further analysis based on changes in the patient's condition. As decisions are made and outcomes ensue, they can then be catalogued for the benefit of future patients, physicians, and researchers.

Given the widespread interest in big data applications in healthcare, the future of this field remains bright for those beginning their careers. Continued progress will be contingent on the influx of talent into the field along with progress in analytics platforms and research methodologies. Funding will likely continue to come from a variety of sources. Currently grants and venture capital are the most significant, but these will be supplemented and perhaps replaced by internal budgets as big data proves its value to the patient, physician, and hospital system. ●

REFERENCES

Anglemyer, A., H.T. Horvath, and L. Bero [2014]. "Healthcare Outcomes Assess with Observational Study Designs Compared with Those Assessed in Randomized Trials," *Cochrane Database Systems Review*, 29(4): DOI: 10.1002/14651858.MR000034.pub2.

Battelle Technology Partnership Practice, for the Pharmaceutical Research and Manufacturers of America (PhRMA) [2015]. "Biopharmaceutical Industry-Sponsored Clinical Trials: Impact on State Economies," March. http://www.phrma.org/sites/default/files/pdf/biopharmaceutical-industry-sponsored-clinical-trials-impact-on-state-economies.pdf.

Frankovich, J., C.A. Longhurst, and S.M. Sutherland [2011]. "Evidence-Based Medicine in the EMR Era," *New England Journal of Medicine*, 365, pp. 1758–1759.

Graham, D.J., D. Campen, R. Hui, et al [2005]. "Risk of Acute Myocardial Infarction and Sudden Cardiac Death in Patients Treated with Cyclo-Oxygenase 2 Selective and Non-Selective Non-Steroid Anti-Inflammatory Drugs: Nested Case Control Study," *Lancet*, 365(9458), pp. 475–481.

Ioannidis, J.P., A.B. Haidich, M. Pappa, et al [2001]. "Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies," *Journal of the American Medical Association*, 286(7), pp. 821–830.