Harvard Medical Student Review Vol. 2 (Issue 1), pp. 24-27.

Article

Preparing a New Generation of Clinicians for the Era of Big Data

Ari Moskowitz MD,[1] Jakob McSparron MD,[1] David J. Stone MD,[3] and Leo Anthony Celi MD, MPH, MS[1,2,*]
[1]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA
[2]Laboratory of Computational Physiology, Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Departments of Anesthesiology and Neurological Surgery, and the Center for Wireless Health, University of Virginia, Charlottesville, VA 22908, USA

*Correspondence: lceli@bidmc.harvard.edu

Synopsis: As medicine becomes increasingly complex and financially constrained, it will be the responsibility of every clinician to understand and participate in the enterprise of extracting lessons learned from digitally captured patient care.

Introduction

In the past, both outpatient and inpatient clinical data were entered and stored in paper formats that were not systematically organized, were accessible only at a single physical place at a time, and were stored—when not lost—in distant, variably efficient medical records departments. With the current, widespread adoption of Electronic Health Records (EHRs), such data should now be made available and leveraged to generate knowledge. But even post-digitization, the information generated from everyday patient encounters remains under-utilized. Our ability and capacity to train both new and experienced clinicians to manage this tremendous amount of data lag far behind the pace of the data revolution. Medical education at all levels must come to address data management and utilization issues as we enter the era of Big Data in the clinical domain.

In this paper, we review the potential that Big Data holds in knowledge discovery in medicine and propose to incorporate more data science into the medical school curriculum. In this context, we present our work with the Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC) database, including the Critical Data marathons that we organize.

The Problem: Data Deserts and Dead Ends

Diagnostic and therapeutic technologies continue to evolve rapidly, and both individual practitioners and clinical teams face increasingly complex decisions. Unfortunately, the current state of medical knowledge does not provide the guidance to make the majority of clinical decisions on the basis of evidence: According to the 2012 Institute of Medicine Committee Report, only 10%–20% of clinical decisions are evidence based. The problem even extends to the creation of clinical practice guidelines (CPGs). Nearly 50% of recommendations made in specialty society guidelines rely on expert opinion rather than experimental data (Committee on the Learning Health Care System in America, 2012; Kung et al., 2012). Furthermore, the creation process of CPGs is "marred by weak methods and financial conflicts of interest," rendering current CPGs potentially less trustworthy (Steinbrook, 2014).

The present research infrastructure is inefficient and frequently produces unreliable results that cannot be replicated (Steinbrook, 2014). Even randomized controlled trials (RCTs), the traditional gold standards of the research reliability hierarchy, are not without limitations. They can be costly, labor intensive, and slow and can return results that are seldom generalizable to every patient population. It is impossible for a tightly controlled RCT to capture the full, interactive, and contextual details of the actual issues that arise in real clinics and inpatient units. Furthermore, many pertinent but unresolved clinical and medical systems issues do not seem to have attracted the interest of the research enterprise, which has come to focus instead on cellular and molecular investigations and single-agent (e.g., a drug or device) effects. For clinicians, the end result is a bit of a "data desert" when it comes to making decisions.

Electronic medical record (EMR) data are digitally archived and can subsequently be extracted and analyzed. Between 2011 and 2019, the prevalence of EMRs is expected to grow from 34% to 90% among office-based practices, and the majority of hospitals have replaced or are in the process of replacing paper systems with comprehensive, enterprise EMRs (Committee on the Learning Health Care System in America; Hsiao et al., 2011; Macleod et al., 2014). The power of scale

intrinsic to this digital transformation opens the door to a massive amount of currently untapped information. The data, if properly analyzed and meaningfully interpreted, could vastly improve our conception and development of best practices. The possibilities for quality improvement, increased safety, process optimization, and personalization of clinical decisions range from impressive to revolutionary. The National Institutes of Health (NIH) and other major grant organizations have begun to recognize the power of Big Data in knowledge creation and are offering grants to support investigators in this area (http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-13-009.html).

Already, a number of organizations and academic medical centers have begun to harness the potential of Big Data through application in both clinical and research arenas. To cite a few examples, the Mayo Clinic has implemented software that were developed using clinical data, including the Ambient Warning and Response Evaluation (AWARE) system that supports best practice in the ICU and operating room; Syndromic Surveillance, which provides "sniffers" to detect sepsis; and YES Board, a multi-patient management tool that offers real-time situational awareness for the Emergency Department (Milliard, 2014). At Cleveland Clinic, medical calculators have been developed that take into account patient demographics as well as details about the medical condition in order to guide clinicians and patients in decision making with regard to tests and treatments (Landro, 2014). Finally, efforts are underway to build international clinical databases. With funding from the NIH, the Laboratory of Computational Physiology at the Harvard-MIT Division of Health Science and Technology is spearheading an initiative to create an open-access repository of EHR data from ICUs across partner countries, including the US, Belgium, the United Kingdom, and France (Celi et al., 2013). Funded by the European Commission, the Brain Monitoring with Information Technology (BrainIT) group has created a core data set collected from 20 neurointensive care centers from 11 countries across Europe (http://www.brain-it.eu/).

This digital transformation has taken place before the eyes, but below the radar screen, of the medical education system. While clinical workflow has been irreversibly altered by the implementation of these systems, education and training largely proceed without taking these issues into account. Unfortunately, few physician educators are adequately trained in data management and analysis (Lucey, 2013). Little to no time in training is spent teaching the fundamentals of data science, knowledge creation, and outcomes-based practice. Most medical schools devote a single month to basic epidemiology and statistics (Looney et al., 1998; https://www.aamc.org/initiatives/medaps/curriculumreports/). Residency curriculums require physicians-in-training to utilize a variety of software applications for patient care, but there are few resources dedicated to improve our use of the information we create: the general systems principles involved in the digital transformation of medicine are not being conveyed, perhaps because they are still in the process of being formulated.

The Approach: Leveraging Data to Meet Clinical Needs

Beyond simple user principles, trainees do not learn the skills and concepts necessary for the optimal use of EMRs, including knowledge creation and personalized clinical decision making through analysis of large data sets. To date, this is largely because such systems have not been designed or implemented with these goals in mind. In the coming era of "Big Data," our community of medical educators and researchers must leverage digital systems for this purpose and find a way to prepare trainees for this critical role. Most current medical educators are not particularly well versed in these issues, which arose after their own training and represent distinctive areas of knowledge lying outside the historical clinical domain. Academic medical informatics departments should be actively enlisted and involved in the response to this challenge. It is likely that medical educators themselves will need to be educated by internal and external experts before they can proceed to educate others. Maximizing knowledge generation from EMRs will require some redefinition of the roles of the contemporary doctor. Many barriers exist to incorporating new courses into already overloaded medical school curricula. A reappraisal is needed to determine what can be omitted or taught more efficiently. As an example, elective rotations in biostatistics and Big Data could be offered to fourth year medical students. The fourth year medical school curriculum is generally less structured than the first three years (Walling and Merando, 2010), and an introduction to secondary use of EHR data may provide a foundation for students to be able to contribute to knowledge discovery regardless of the career path they eventually choose.

In order to support this transformation in clinical practice and research, physicians-in-training will need to be educated to some reasonable degree in the analysis of large data sets in collaboration with data scientists and biostatisticians. The multi-disciplinary team is now expanding beyond nurses, pharmacists, and other traditional allied health personnel and will include individuals with advanced data analytic abilities.

Our group has been working with data scientists from the Massachusetts Institute of Technology (MIT) and biostatisticians from the Harvard School of Public Health using the MIMIC database. This database, which holds clinical data from over 60,000 stays at the intensive care units (ICUs) at Beth Israel Deaconess

Medical Center, has been meticulously de-identified and is freely shared online with the research community (Saeed et al., 2011). It provides a platform from which "crowdsourcing" can be applied in the generation of hypotheses, discovery of knowledge, and evidence creation in the practice of critical care.

MIMIC is a public-access database, and our group actively encourages participation from clinicians at all levels of training, including medical students, residents, fellows, and faculty. Clinicians are partnered with data scientists from the Massachusetts Institute of Technology and the Harvard School of Public Health. The clinician-data scientist team, under the supervision of an expert in the field of clinical informatics, extracts data from MIMIC and preforms the necessary analyses to answer questions that arise during rounds.

Current Work and Outcomes

To date, more than 50 clinicians, including doctors, nurses, and pharmacists, have worked alongside data scientists on a wide range of projects. These projects have already begun to answer the kinds of novel clinical questions that would have taken far more time and resources to address in a RCT. Through the use of data extracted from the MIMIC database, many original research articles have already been published, and many more are in the pipeline. The range of topics is broad and includes studies exploring the optimal dosing of medications, the creation of prediction models, and the discovery of previously unknown or underappreciated relationships. To date, over 50 journal articles have been published using data from MIMIC (https://mimic.physionet.org/about/publications.html).

Importantly, medical students and resident clinicians frequently maintain their relationship with the data scientists as they progress in their training. The highly accessible and open nature of MIMIC allows for continuation of academic projects even from remote locations. This is producing a cohort of physicians that is both cognizant and capable of dealing with the data issues that arise from digitalization and its application to clinical process and outcome improvements.

In January 2014, our group hosted the Critical Data Marathon and Conference. In the Hackathon, physicians, nurses, and pharmacists were paired with data scientists and encouraged to investigate a variety of clinical questions that arise in the ICU. Over a 2-day period, over 150 attendees began to answer questions such as whether acetaminophen should be used to control fevers in critically ill patients and what the optimal mean arterial pressure is among septic patients. This event fostered relationships between clinicians and data scientists that will support ongoing research in the ICU setting. We currently plan for this conference to be a regular event and would encourage similar endeavors at other institutions.

Overall, the key contribution of the MIMIC database and the Critical Data Marathons is the promotion of ongoing, cross-disciplinary collaboration around learning. Clinicians, including nurses and pharmacists, are provided a platform to contribute to knowledge discovery that had been traditionally exclusive to academic researchers. Data scientists are thrilled by the opportunity to transform practice and improve health outcomes. Creating and fostering these partnerships across disciplines is non-trivial, given that their paradigms and practices are difficult to interlace into new and different contexts. However, the Critical Data Marathons simultaneously held at MIT, in London, and in Paris in September 2014 proved that this cultural shift is not only feasible but also replicable and scalable.

Next Steps/Discussion

We envision a learning system where knowledge generation is routine and fully integrated into the clinical workflow. The next step toward this goal is the introduction of formal courses into medical school and residency curricula. These courses will focus on skills needed to build, maintain, and analyze large data sets. The growing importance of Big Data in everyday clinical situations will be emphasized, and students will be given the opportunity to investigate clinical questions that have come up in the course of their clinical rotations.

In addition to the above curricular goals, medical schools and residency training programs need to develop new ways of incorporating data scientists into a multi-disciplinary approach to patient care. As Big Data becomes more accessible, individuals who can navigate and help analyze large data sets will become an increasingly important part of the care team. Physicians will need to know how to work with these professionals in ways that allow for meaningful conclusions to be drawn from large amounts of data in real time. Beyond partnerships formed in future Critical Data Marathons, we envision a program where data scientists from the MIMIC group will be able to join clinical teams for ICU rounds and participate in clinical decision making through the real-time analysis of Big Data.

The inclusion of data scientists as part of a multi-disciplinary team is one way to engender collaboration with the clinicians. This model has been successful with other healthcare professions, including pharmacy, physical therapy, and social work. Participating in rounds will provide data scientists an opportunity to work alongside clinicians and to receive immediate feedback to their input. Their addition to the team creates a substrate for potential innovation chain reactions that simply would not occur under other circumstances.

The time has come to leverage the data we generate during routine patient care to formulate a more

complete lexicon of evidence-based recommendations and support shared decision making with our patients. In this setting, the practice of every clinician will necessarily expand to participate in these issues. This must be done without creating enormous extra work for already overburdened clinicians. In contrast, the interactions of regular clinicians with a knowledge-generating system should be rewarding in both the intellectual and workflow senses. All clinicians will share the responsibility of creating a more complete knowledge base and transforming practice to improve care. The data are already being generated. Now is the time to train clinicians who can harness the true potential of this information to provide better care for our patients.

Acknowledgments

L.A.C. is funded by the National Institute of Health R01 grant R01 EB017205-01A1.

References

Celi, L.A., Mark, R.G., Stone, D.J., and Montgomery, R.A. (2013). "Big data" in the intensive care unit. Closing the data loop. Am. J. Respir. Crit. Care Med. *187*, 1157–1160. http://dx.doi.org/10.1164/rccm.201212-2311ED.

Committee on the Learning Health Care System in America (2012). Institute of Medicine, Best Care at Lower Cost: The Path to Continuously Learning Health Care in America (Washington, DC: The National Academies Press).

Hsiao, C.-J., Hing, E., Socey, T.C., and Cai, B. (2011). Electronic health record systems and intent to apply for meaningful use incentives among office-based physician practices: United States, 2001-2011 (Hyattsville, MD: National Center for Health Statistics).

Kung, J., Miller, R.R., and Mackowiak, P.A. (2012). Failure of clinical practice guidelines to meet institute of medicine standards: Two more decades of little, if any, progress. Arch. Intern. Med. *172*, 1628–1633. http://dx.doi.org/10.1001/2013.jamainternmed.56.

Landro, L. (2014) Medical Calculators use Big Data to Help Patients Make Choices. The Wall Street Journal, September 14, 2014. http://online.wsj.com/articles/medical-calculators-use-big-data-to-help-patients-make-choices-1410724818.

Looney, S.W., Grady, C.S., and Steiner, R.P. (1998). An update on biostatistics requirements in U.S. medical schools. Acad. Med. *73*, 92–94.

Lucey, C.R. (2013). Medical education: part of the problem and part of the solution. JAMA Intern. Med. *173*, 1639–1643. http://dx.doi.org/10.1001/jamainternmed.2013.9074.

Macleod, M.R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J.P., Al-Shahi Salman, R., Chan, A.W., and Glasziou, P. (2014). Biomedical research: increasing value, reducing waste. Lancet *383*, 101–104. http://dx.doi.org/10.1016/S0140-6736(13)62329-6.

Milliard, M. (2014) Mayo Clinic Launches Bedside Analytics. Healthcare IT News, March 20, 2014. http://www.healthcareitnews.com/news/mayo-clinic-launches-bedside-analytics.

Saeed, M., Villarroel, M., Reisner, A.T., Clifford, G., Lehman, L.W., Moody, G., Heldt, T., Kyaw, T.H., Moody, B., and Mark, R.G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. Crit. Care Med. *39*, 952–960. http://dx.doi.org/10.1097/CCM.0b013e31820a92c6.

Steinbrook, R. (2014). Improving clinical practice guidelines. JAMA Intern. Med. *174*, 181. http://dx.doi.org/10.1001/jamainternmed.2013.7662.

Walling, A., and Merando, A. (2010). The fourth year of medical education: a literature review. Acad. Med. *85*, 1698–1704. http://dx.doi.org/10.1097/ACM.0b013e3181f52dc6.