Big data in healthcare– the promises, challenges and opportunities from a research perspective: A case study with a model database

Mohammad Adibuzzaman, PhD[1], Poching DeLaurentis, PhD[1], Jennifer Hill, MSc[1], Brian D. Benneyworth, MD, MS[2]
[1]Regenstrief Center for Healthcare Engineering, Purdue University, West Lafayette, Indiana, USA
[2]Children's Health Services Research Group, Department of Pediatrics, Indiana University School of Medicine, Indianapolis, USA

Abstract

Recent advances in data collection during routine health care in the form of Electronic Health Records (EHR), medical device data (e.g., infusion pump informatics, physiological monitoring data, and insurance claims data, among others, as well as biological and experimental data, have created tremendous opportunities for biological discoveries for clinical application. However, even with all the advancement in technologies and their promises for discoveries, very few research findings have been translated to clinical knowledge, or more importantly, to clinical practice. In this paper, we identify and present the initial work addressing the relevant challenges in three broad categories: data, accessibility, and translation. These issues are discussed in the context of a widely used detailed database from an intensive care unit, Medical Information Mart for Intensive Care (MIMIC III) database.

1 Introduction

The promise of big data has brought great hope in health care research for drug discovery, treatment innovation, personalized medicine, and optimal patient care that can reduce cost and improve patient outcomes. Billions of dollars have been invested to capture large amounts of data outlined in big initiatives that are often isolated. The National Institutes of Health (NIH) recently announced the *All of Us* initiative, previously known as the *Precision Medicine Cohort Program*, which aims to collect one million or more patients' data such as EHR, genomic, imaging, socio-behavioral, and environmental data over the next few years[1]. *The Continuously Learning Healthcare System* is also being advocated by the Institute of Medicine to close the gap between scientific discovery, patient and clinician engagement, and clinical practice[2]. However, the big data promise has not yet been realized to its potential as the mere availability of the data does not translate into knowledge or clinical practice. Moreover, due to the variation in data complexity and structures, unavailability of computational technologies, and concerns of sharing private patient data, few projects of large clinical data sets are made available to researchers in general. We have identified several key issues in facilitating and accelerating data driven translational clinical research and clinical practice. We will discuss in-depth in the domains of data quality, accessibility, and translation. Several use cases will be used to demonstrate the issues with the "Medical Information Mart for Intensive Care (MIMIC III)" database, one of the very few databases with granular and continuously monitored data of thousands of patients[3].

2 Promises

In the era of genomics, the volume of data being captured from biological experiments and routine health care procedures is growing at an unprecedented pace[4]. This data trove has brought new promises for discovery in health care research and breakthrough treatments as well as new challenges in technology, management, and dissemination of knowledge. Multiple initiatives were taken to build specific systems in addressing the need for analysis of different types of data, e.g., integrated electronic health record (EHR)[5], genomics-EHR[6], genomics-connectomes[7], insurance claims data, etc. These big data systems have shown potential for making fundamental changes in care delivery and discovery of treatments such as reducing health care costs, reducing number of hospital re-admissions, targeted interventions for reducing emergency department (ED) visits, triage of patients in ED, preventing adverse drug effects, and many more[8]. However, to realize these promises, the health care community must overcome some core technological and organizational challenges.

3 Challenges

3.1 Data

Big data is not as big as it seems

In the previous decade, federal funding agencies and private enterprises have taken initiatives for large scale data collection during routine health care and experimental research[5,9]. One prominent example of data collection during routine health care is the Medical Information Mart for Intensive Care (MIMIC III) which has collected data for more than fifty thousand patients from Beth Israel Deaconess Hospital dating back to 2001[3]. This is the largest publicly available patient care data set of an intensive care unit (ICU) and an important resource for clinical research. However, when it comes to identifying a cohort in the MIMIC data for answering a specific clinical question, it often results in a very small set of cases (small cohort) that makes it almost impossible to answer the question with a strong statistical confidence. For example, when studying the adverse effects of a drug-drug interaction, a researcher might be interested in looking at the vital signs and other patient characteristics during the time two different drugs were administered simultaneously, including a few days before the combination and a few days after the combination. Often this selection criteria results in a very small cohort of patients limiting the interpretation of the finding and with statistically inconclusive results. As an example, a researcher may want to investigate if any adverse effect exists when anti-depressants and anti-histamines are administered simultaneously. A query of simultaneous prescriptions of Amitriptyline HCl (anti-depressant) and Diphenhydramine HCl (anti-histamines) returned only 44 subjects in the MIMIC database (Figure 1). Furthermore, by filtering the data with another selection criterion (e.g., to identify the subjects for which at least one day's worth of data exist during, before and after the overlap) the query returned a much smaller cohort with only four records.

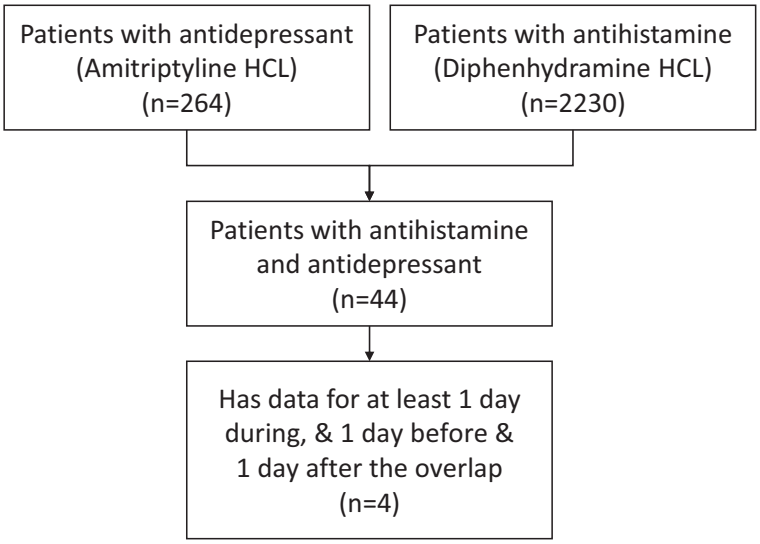

Figure 1: Example of a small cohort with clinical selection criteria.

Data do not fully capture temporal and process information

In most cases, clinical data are captured in various systems, even within an organization, each with a somewhat different intent and often not well integrated. For example, an EHR is primarily used for documenting patient care and was designed to facilitate insurance company billing[10], and pharmacy records were designed for inventory management. These systems were not developed to capture the temporal and process information which is indispensable for understanding disease progression, therapeutic effectiveness and patient outcomes. In an attempt to study clinical process of vancomycin therapeutic drug monitoring based on ICU patient records in the MIMIC database, it was discovered that such process is not easy to reconstruct. Ideally, a complete therapeutic process with a particular drug contains the history of the drug's prescription, each of its exact administration times, amount and rate, and the timing and measurements of the drug in the blood throughout the therapy. From the MIMIC III database we were able to find prescription

information but it lacks the detailed dosing amount and prescription's length of validity. The "inputevents" table contains drug administration information but does not include the exact time-stamp and drug amount which is critical for studying intravenous infused vancomycin in the ICU . It is also difficult to match drug prescription and administration records because their recording times in the clinical systems often are not the precise event times, and prescribed drugs are not always administered.

Time	Prescriptions						Input_events					Lab_events		
start_time/chart_time	end_date	drug	prod_strength	dose_val_rx	dose_unit_rx	route	endtime	label	amount	amountuom	rate	value	valueuom	label
10/13/2141 0:00	10/16/2141 0:00	Vancomycin	1g Frozen Bag	1000	mg	IV								
10/13/2141 8:30							10/13/2141 8:31	Vancomycin	1	dose				
10/13/2141 18:01												33.4	ug/mL	VANCOMYCIN
10/14/2141 8:31												22.4	ug/mL	VANCOMYCIN
10/14/2141 19:50												17.1	ug/mL	VANCOMYCIN
10/14/2141 22:00							10/14/2141 22:01	Vancomycin	1	dose				
10/15/2141 6:18												23.1	ug/mL	VANCOMYCIN
10/15/2141 19:50												3.8	ug/mL	VANCOMYCIN
10/15/2141 21:10							10/15/2141 21:11	Vancomycin	1	dose				

Figure 2: An example of vancomycin therapeutic process reconstruction of one unique ICU stay using data from three different tables in the MIMIC III database.

Moreover, since the MIMIC III database does not contain detailed infusion event records which may be available from infusion pump software, one cannot know the precise drug infusion amount (and over what time) for any particular administration. The sparse and insufficient information on drug administration makes it almost impossible to associate available laboratory records and to reconstruct a therapeutic process for outcomes studies. Figure 2 is such an attempt of process reconstruction using data from the MIMIC III database including prescriptions, input events, and lab events for one patient during a unique ICU stay. The record only shows one valid prescription of vancomycin for this patient with start and end dates but does not indicate the administration frequency (e.g., every 12 hours) or method (e.g., continuous or bolus). The input events data (the second main column) came from the nursing records but it only shows one dose of vancomycin administration on each of the three-day ICU stay: one in the morning and two in the evening. Even though, as shown in the third main column, the "lab event" data contain the patient's vancomycin concentration levels measured during this period, without the exact amount and duration of each vancomycin infusion, it is difficult to reconstruct this particular therapeutic process for the purposes of understanding its real effectiveness.

The problem of missing data remains relevant, even when the nursing workflow was designed to capture the data in the EHR. For example, as part of the nursing workflow, the information of drug administration should be documented in the medication administration records each time vancomycin was administered, and the MIMIC system was designed to capture all. But this was often not the case from our review of the database 2. Additionally, often times a patient's diagnoses, co-morbidities, and complications are not fully captured nor available for reconstructing the complete clinical encounter. Those pieces of information are usually documented as free text not discrete data that can easily be extracted. Moreover, precise timings of the onset of an event and its resolution are rarely present. In the previous example of analyzing the effect of simultaneously administering Amitriptyline HCl and Diphenhydramine HCl, based on our selection criteria, we were able to find only one or two cases where such data were recorded (Figure 3). In the figure, each color represents one subject, and only one color (green, ID:13852) is consistently present in the time window for the selection criteria indicating missing systolic blood pressure measurements for the other subjects. This example is not an exception for cohort selection from data captured during care delivery, but a common occurrence[11], due to the complex nature of the care delivery process and technological barriers in the various clinical systems developed in the past decade or so.

3.2 Access

Accessibility to patient data for scientific research and sharing of the scientific work as digital objects for validation and reproducibility is another challenging domain due to patient privacy concerns, technological issues such as interoperability, and data ownership confusion. This has been a widely discussed issue in recent years of the so-called patient or health data conundrum as individuals do not have easy access to their own data[12]. We are discussing these challenges in the context of privacy, share-ability, and proprietary rights as follows.

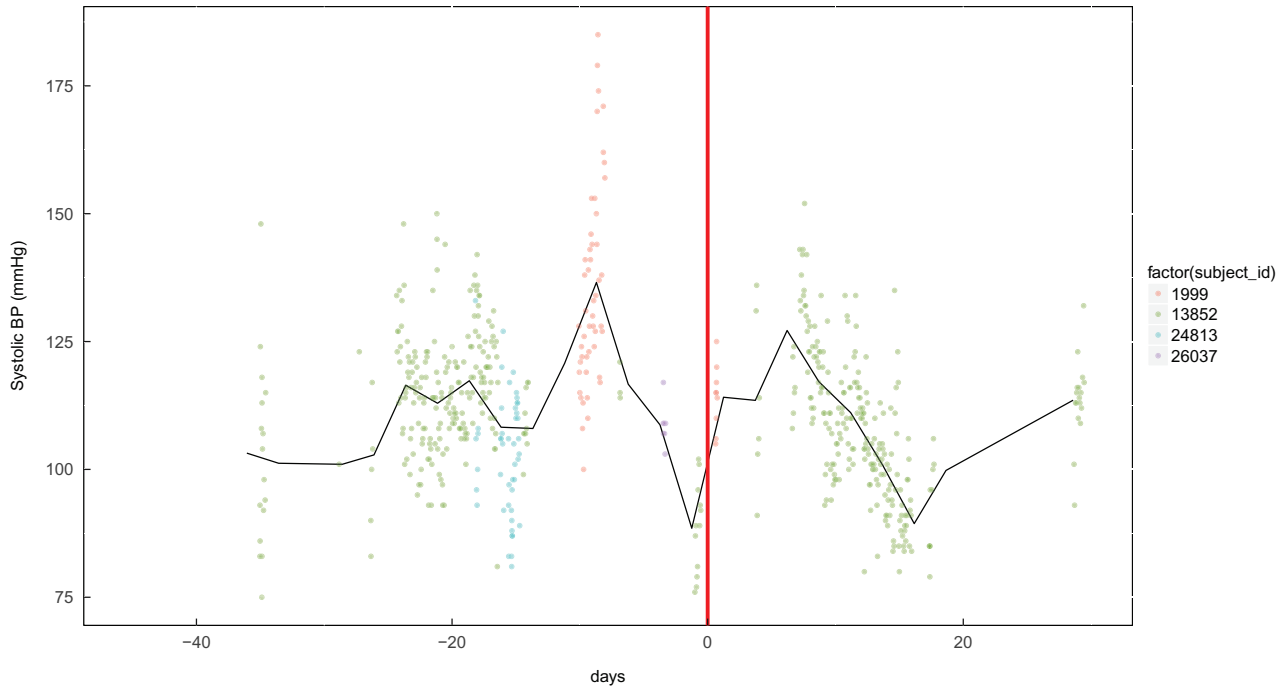

Figure 3: Example of a cohort with missing systolic blood pressure data for three out of the four subjects meeting our clinical selection criteria. Day 0 (zero) is when drug overlap begins. This start of the overlap is aligned with multiple subjects and is denoted by the thick red line. Each data point represents one measurement from the "chartevents" table and each color indicates one subject and the black line indicates the average of the selected cohort.

Privacy

Access to health care data is plagued by vulnerability due to patient privacy considerations which are protected by federal and local laws of protected health information such as Health Insurance Portability and Accountability Act of 1996 (HIPAA)[13]. The fear of litigation and breach of privacy discourages providers from sharing patient health data, even when they are de-identified. One reason is that current approaches to protect private information is limited to de-identification of an individual subject with an ID, which is vulnerable to twenty questions-like problems. For example, a query to find any patient who is of Indian origin and has some specific cancer diagnosis with a residential zip code 3-digit prefix '479' may result in only one subject; thus exposing the identity of the individual.

Share-ability

Even after de-identification of patient data, the sharing of such data and research works based on the data is a complicated process. As an example, "Informatics for Integrating Biology and the Bedside (i2b2)"[5] is a system designed to capture data for scientific research during routine health care. i2b2 is a large initiative undertaken by Partners Healthcare System as an NIH-funded National Center for Biomedical Computing (NCBC). It contains a collection of data systems with over 100 hospitals that are using this software system on top of their clinical database. As a member of this project, each participating hospital system needed to transform their data into a SQL based star schema after de-identification. It required much effort for each institution to make the data available for scientific research as well as to develop the software in the first place. Although i2b2 was used exhaustively for research, sharing of data and research work as digital objects (i.e., the coding and the flow of the analysis) is not easily achieved. We argue that current EHR and other clinical systems do not empower the patients to take control of their data and engage in citizen science. The crowd sourcing approach might be one way to make a paradigm shift in this area which, unfortunately, is not yet possible with the current systems such as i2b2. A good example is the success in open source software technologies in other disciplines and applications (such as Linux, Git-hub, etc.) which rely on the engagement of many talented and passionate scientists and engineers all over the world to contribute their working products as digital objects[14].

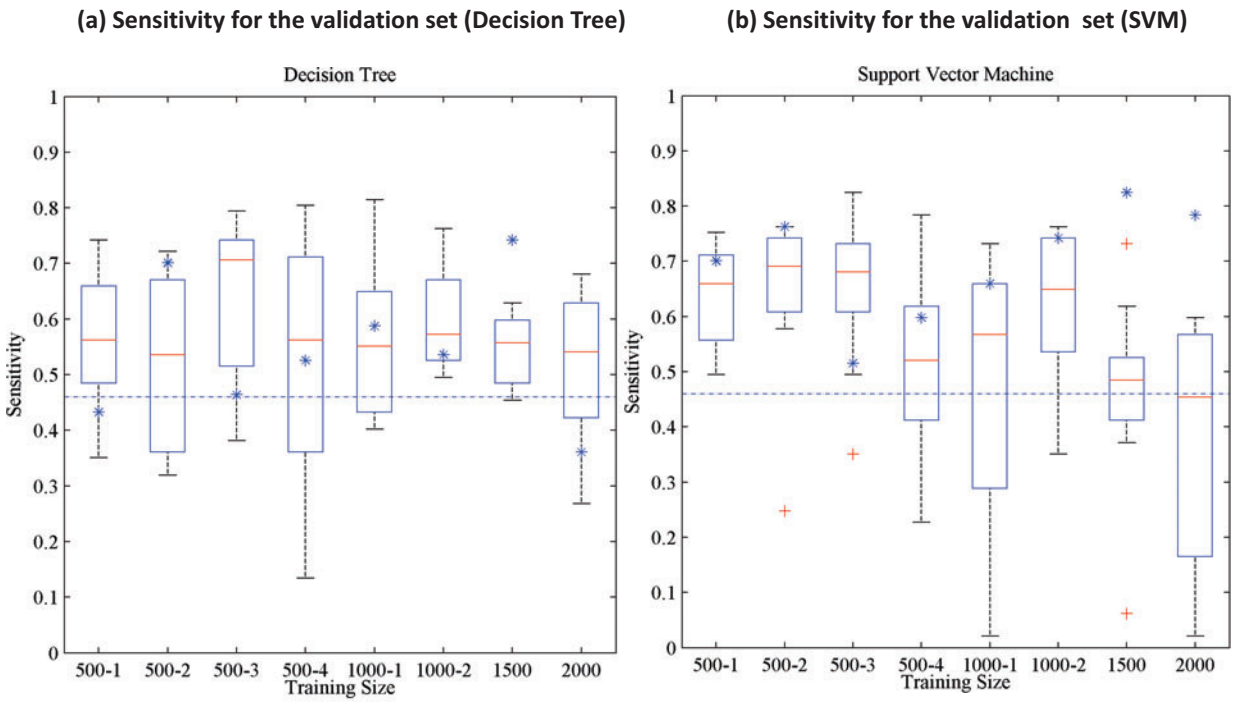

Figure 4: Sensitivity for the machine learning algorithms for different training sizes for prediction of Medical Emergency Team (MET) activation from the MIMIC database[15]. The X-axis represents training size for different trials. For each training set, the results of a 10 fold cross validation are reported as box plots (the central red line is the median, the edges of the box are the 25^{th} and 75^{th} percentiles, the whiskers extend to the extreme data points the algorithm considers to be not outliers, and the red + sign denotes outliers). The blue asterisks represent the performance on the validation set of the algorithm that performs best on the test set. The blue dashed lines represent the performance of the National Early Warning Score (NEWS)[16].

Proprietary rights

A relevant issue is the ongoing debate about the ownership of patient data among various stakeholders in the healthcare system including providers, patients, insurance companies and software vendors. In general, the current model is such that the patient owns his/her data, and the provider stores the data with proprietary software systems. The business models of most traditional EHR companies, such as Epic and Cerner, are based on building proprietary software systems to manage the data for insurance reimbursement and care delivery purposes. Such approach does not encourage or makes it difficult for individual patients to share data for scientific research, nor does it encourage patients to obtain their own health records that may help better manage their care and improve patient engagement.

3.3 Translation

Historically, a change in clinical practice is hard to achieve because of the sensitivity and risk aversion of care delivery. As an example, the use of beta blockers to prevent heart failure took 25 years to reach a widespread clinical adoption after the first research results were published[2]. This problem is much bigger for big data driven research findings to be translated into clinical practice because of the poor understanding of the risks and benefits of data driven decision support systems. Many machine learning algorithms work as a "black box" with no provision of good interpretations and clinical context of the outcomes, even though they often perform with reasonable accuracy. Without proper understanding and translatable mechanisms, it is difficult to estimate the risk and benefit of such algorithms in the clinical setting and thus discourages the new methods and treatments from being adopted by clinicians or approved by the regulatory bodies such as the FDA.

For example, if a machine learning algorithm can predict circulatory shock from patient arterial blood pressure data, what would be the risk if the algorithm fails in a particular setting based on patient demographics or clinical history? What should be the sample size to achieve high confidence in the results generated by the algorithm? These are some critical questions that cannot be answered by those traditional "black box" algorithms, nor have they been well accepted by the medical community, which relies heavily upon rule based approaches.

As an example, a decision tree algorithm might perform very differently for prediction of Medical Emergency Team (MET) activation based on the training set or sample size from the MIMIC data. Furthermore, the prediction result can be very different when another machine learning algorithm, the support vector machine (SVM), was used (Figure 4).

3.4 Incentive

Yet another barrier in using big data for better health is the lack of incentive for organizations to take initiative to address the technological challenges. As mentioned earlier, EHRs are developed for purposes other than knowledge advancement or care quality improvement, and that has led to unorganized, missing, and inadequate data for clinical research. An individual health system does not usually have the incentive to make these data organized and available for research, unless they are big academic institutions. It would be easier for each individual health system to share data if they were organized and captured using standard nomenclature and with meaningful and useful detailed information with significant detail. A key question any health organization faces is: what is the return on investment for my hospital to organize all the clinical data it gathers? One model is the Health Information Technology for Economic and Clinical Health Act (HITECH) which promotes the adoption and meaningful use of health information technology. The act authorized incentive payments be made through Medicare and Medicaid to clinicians and hospitals that adopted and demonstrated meaningful use of EHRs, and the US government has committed payments up to $27 billion dollars over a ten year period[17]. This incentive has paved the way for widespread adoption of EHRs since HITECH was enacted as part of the American Recovery and Reinvestment Act in 2009. However, for the purpose of using clinical data for scientific innovation and improving care delivery process, no apparent financial incentives currently exist for any organization to do so.

4 Opportunities

4.1 Data

For data driven research in health care, we propose to record the most granular data during any care delivery process so as to capture the temporal and process information for treatment and outcomes. For example, in an intensive care unit, the exact time of medication administrations need to be captured. This can be achieved in a number of ways. As a nurse bar code scans an oral medication into the electronic medication administration record (eMAR) the system also timestamps the action in the EHR. Detailed intravenous drug infusions can be linked to the patient clinical records by integrating the smart infusion pumps with the EHR systems. The Regenstrief National Center for Medical Device Informatics (REMEDI), formerly known as the Infusion Pump Informatics[18], has been capturing for capturing process and temporal infusion information. The planned expansion of such data set will allow linked patient outcomes and drug admin data forming a more complete treatment process for answering research and treatment effectiveness questions related to the administration of drugs such as drug-drug interaction, safe and effective dosage of drugs, etc., among others.

In order to achieve a statistically significant sample size after cohort selection, we promote breaking the silos of individual clinical data systems and making them interoperable across vendors, types and institutional boundaries with minimal effort. For the next generation of EHRs, these capabilities need to be considered.

4.2 Access

Patient/citizen powered research

To replicate the success in open source technologies in other disciplines by enabling citizen science, data and research analysis must be accessible to everyone. At the same time, patient privacy needs to be protected complying with the

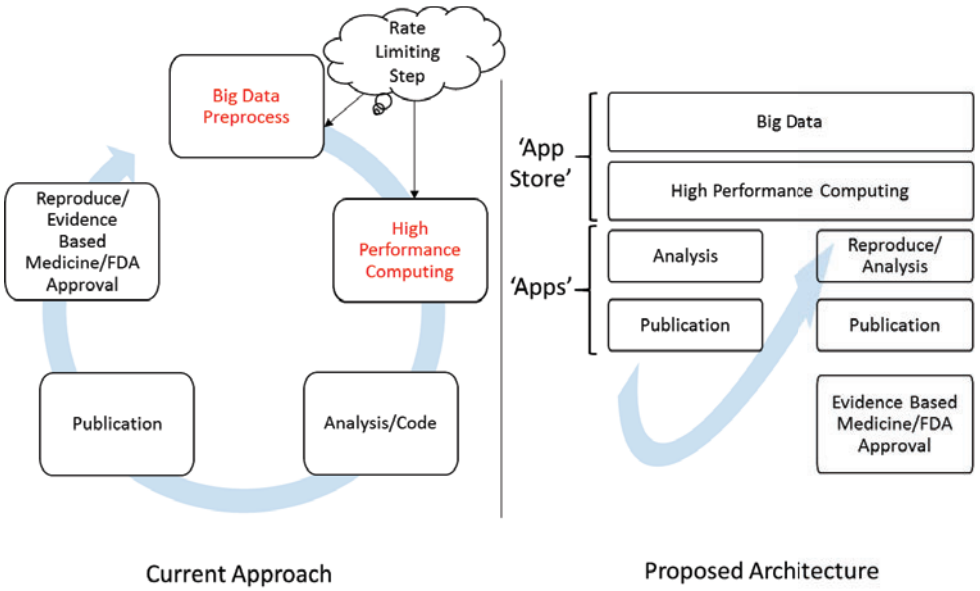

Figure 5: System concept for community driven software-hardware eco-system analogous to 'app store' for data driven clinical research.

privacy law and proprietary rights of the vendors, and researchers need to be protected. As an example, we have demonstrated such a system with the MIMIC database where interoperable and extensible database technologies have been used on de-identified patient data in a high performance computing environment[19].

Shareable digital objects

For the next generation of EHRs and other big data systems such as REMEDI[18] and i2b2[5], data must be findable, accessible, interoperable and reproducible (FAIR)[20]. For big data systems, a software-hardware ecosystem could work as a distribution platform with characteristics analogous to an Apple or Android "app store" where any qualified individual can access the de-identified data with proper authentication without the need for a high throughput infrastructure and the rigorous work, including pre-processing of the data needed to reproduce previous works. The proposed architecture is shown in Figure 5[19].

4.3 Translation

Causal understanding

Historically, clinical problems and treatment are studied and understood as "cause and effect". For example, genetic disposition and lifestyle could lead to frequent urination, fatigue and hunger, and can be associated with diabetes. Based on this, the patient may be treated for this disease. However, most machine learning algorithms do not provide such a rule based approach; rather they predict the outcome of a given set of inputs, which may or may not be associated with known clinical understanding. Unlike other disciplines, clinical applications require a causal understanding of data driven research. Hence, most clinical studies start with some hypothesis, that 'A' causes 'B'. The gold standard to identify this causation is randomized controlled trials (RCTs), which have also been the gold standard for regulatory approval of new drugs. Unfortunately, EHRs and the like data captured during routine healthcare has sampling selection bias and confounding variables and hence it is important to understand the limitation of such data sets. To answer the causal questions, a new generation of methods are necessary to understand the causal flow of treatment, outcome, and molecular properties of drugs by integrating big data systems for analysis and validation of hypothesis for transportability across studies with observational data[21,22]. These methods would enable the regulators to understand the risk and benefit of data driven systems in clinical settings for new guidelines enabling the translation. Once those guidelines are established, technological solution must also be enabled at the point of care such that clinicians

can access for data driven queries as part of their clinical workflow.

5 Conclusion

"Big data" started with many believable promises in health care, but unfortunately, clinical science is different from other disciplines with additional constraints of data quality, privacy, and regulatory policies. We discussed these concepts in pursuit of a holistic solution that enables data driven findings to be translated in health care, from bench to bedside. We argue that the existing big data systems are still in their infancy, and without addressing these fundamental issues the health care big data may not achieve its full potential. We conclude that to make it to the next level, we need a larger cohort of institutions to share more complete, precise, and time stamped data as well as with greater willingness to invest in technologies for de-identifying private patient data for it to be shared broadly for scientific research. At the same time, as more and more "big data" systems are developed, the scientific and regulatory communities need to figure out new ways of understanding causal relationship from data captured during routine health care, that would complement current gold standard methods such as RCTs as well as identify the relationship between clinical practice and outcomes, as there is a wide disparity in the quality of care across the country[2].

References

1. National Institute of Health. https://www.nih.gov/research-training/allofus-research-program, 2017.
2. J Michael McGinnis, Leigh Stuckhardt, Robert Saunders, Mark Smith, et al. *Best care at lower cost: the path to continuously learning health care in America*. National Academies Press, 2013.
3. Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.
4. Michael Eisenstein. Big data: The power of petabytes. *Nature*, 527(7576):S2–S4, 2015.
5. Shawn N Murphy and Adam Wilcox. Mission and sustainability of informatics for integrating biology and the bedside (i2b2). *eGEMs (Generating Evidence Methods to improve patient outcomes)*, 2(2):7, 2014.
6. Jennifer L Hall, John J Ryan, Bruce E Bray, Candice Brown, David Lanfear, L Kristin Newby, Mary V Relling, Neil J Risch, Dan M Roden, and Stanley Y Shaw. Merging electronic health record data and genomics for cardiovascular research a science advisory from the american heart association. *Circulation: Cardiovascular Genetics*, 9(2):193–202, 2016.
7. David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, and WU-Minn HCP Consortium. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.
8. David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7):1123–1131, 2014.
9. Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Blent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, and Erik Larsson. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–404, 2012.
10. Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5):291–304, 2008.
11. Brian J Wells, Amy S Nowacki, Kevin Chagin, and Michael W Kattan. Strategies for handling missing data in electronic health record derived data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 1(3):7, 2013.
12. The New York Times. https://www.nytimes.com/2017/01/02/opinion/the-health-data-conundrum.html, 2017.

13. Health insurance portability and accountability act of 1996.

14. Steve Weber. *The success of open source*. Harvard University Press, 2004.

15. Mohammad Adibuzzaman, David G. Strauss, Stephen Merrill, Loriano Galeotti, and Christopher Scully. Evaluation of machine learning algorithms for multi-parameter patient monitoring. *Student Poster Competition at the US FDA*, 2014.

16. B Williams, G Alberti, C Ball, D Bell, R Binks, L Durham, et al. National early warning score (news): Standardising the assessment of acute-illness severity in the nhs. *London: The Royal College of Physicians*, 2012.

17. David Blumenthal and Marilyn Tavenner. The meaningful use regulation for electronic health records. *N Engl J Med*, 2010(363):501–504, 2010.

18. Steven Witz, Natalie R Buening, Ann Christine Catlin, William Malloy, Julie L Kindsfater, Todd Walroth, Alana Washington, and Richard Zink. Using informatics to improve medical device safety and systems thinking. *Biomedical Instrumentation Technology*, 48(s2):38–43, 2014.

19. Mohammad Adibuzzaman, Ken Musselman, Alistair Johnson, Paul Brown, Zachary Pitluk, and Ananth Grama. Closing the data loop: An integrated open access analysis platform for the mimic database. *Computing in Cardiology*, 2016.

20. Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

21. Judea Pearl, Elias Bareinboim, et al. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.

22. Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 540–547. IEEE, 2011.