Statement of Research

Mohammad Adibuzzaman
madibuzz@purdue.edu

January 27, 2020

I want to liberate the power of data that are trapped in siloed systems and disciplines to improve human health. I am a computational scientist by training with extensive experience in model, method and system development for large clinical data analysis and research infrastructure development with high performance computing system. My motivation to work in healthcare arises from a deep desire to contribute in the human well being, live life happily and abundantly. I want to use technology, extremely large data sources, and new methods to complement the clinical expertise for the future of medicine.

1 Experience

I have an undergraduate degree in Computer Science and Engineering. After my undergraduate, I worked for a while as a software engineer, at a software company and also at National University of Singapore. Unsure about the direct impact of a software engineer in human well being, I started my PhD in Computational Science. It is at this time, I was introduced to clinical questions that require sophisticated technological skills such as mathematical modeling, data analysis in super computers, and complement that skill with clinical knowledge. It started with a project to detect hemorrhage from blood pressure data at the US Food and Drug Administration [1], and I never had to think again what I want to do in my life. I realized there is tremendous opportunity to use my skills in medicine. Since then, I have worked in many projects that uses EHR, physiological or sensor data for clinical research as the lead scientist. I have continued this as a Research Scientist at the Regenstrief Center for Healthcare Engineering at Purdue.

2 Achievements

Here at Regenstrief, I have built the research infrastructure for data analysis from the ground up, established partnership with the Laboratory for Computational Physiology led by Roger Mark, and industry partnership for state of the art distributed database technology with Paradigm4, founded by Turing Award recipient MIT Computer Scientist Mike Stonebroker. I have created a new line of research at RCHE for explainable Artificial Intelligence (AI) in healthcare, and actively collaborating with Elias Bareinboim, professor of Computer Science. The goal is to introduce new methods of explainable AI developed by Turing Award recipient Judea Pearl, in clinical research. Built on these foundations, I have secured a grant of $100,000 to develop a cloud computing environment

with HIPAA compliant servers for EHR and physiological data of thousands of patients from Beth Israel Deaconess Hospital. I have also secured a $300,000 data science initiative grant from Purdue for explainable AI for clinical understanding and translation in healthcare. At the same time, I have published many of these works in top journals and conferences [2, 3, 4], and many more are in the pipeline. I have also successfully submitted an NSF grant for reproducibility of data driven research with a budget of $2M, with the application currently under review.

3 Research Plan

3.1 Software-hardware-data ecosystem for collaborative cloud computing framework

I want to develop critical software technologies in support of analytics applications on large integrated data repositories such as electronic health records (EHR), medical devices and biological data. This is based on our initial work of creating a collaborative cloud computing platform with the MIMIC database [2, 4]; we now have fifty researchers actively using the system. Based on this initial work, I aim to develop modular and extensible support for reproducibility of analytics experiments, transportability and generalizability of analytics results, statistical significance and verifiability of outcomes, and demonstration on a core set of analytics kernels. This line of effort will result in production quality cloud-based software for supporting higher level analytics functionality. The software will be integrated into commonly accessed data repositories and analytics frameworks. It will be made available over the public domain, in open source form, as well as in the form of services, accessible over the web. It will support flexible and extensible analytics APIs, on which other systems can be developed and deployed. Underlying this software will be a suite of novel statistical, algorithmic, and distributed computing techniques, with broad applicability.

This research agenda puts forth a principled approach and associated software realizations for four analytics tasks: i) cohort selection - given a defining set of characteristics, identify a sample from among the repository, which serves as a representative for a data study, ii) querying and statistical characterization - given a sequence of queries (statistical as well as traditional database queries), will create composite query objects that capture query sessions, which can be re-instantiated on the same data object, at a later time, iii) support for core analytics tasks - common analytics tasks such as correlation, clustering, association, and classification, must be encapsulated in analytics objects that can be replayed on specified versions of data, and iv) methods for quantifying generalizability and transportability - particularly in the context of sensitive analytics tasks such as causal inference, which are particularly susceptible to cohort selection (see next section).

Effective analyses of emerging EHR and other clinical databases holds the potential for significant benefits in terms of human well-being and socio-economic factors. However, many of the existing systems only provide rudimentary analyses support (primarily querying and aggregate data reporting), and none provide support for generalizability, transportability, verification, validation, and privacy beyond deidentification. To this end, this proposed research effort will provide critical and novel enabling technologies for software systems that are expected to contribute over $400B to the US economy over the next 10 years. A well-designed, modular software system, as proposed, has broad and significant impact across a rich set of cyberinfrastructure.

3.2 Artificial intelligence (AI)/ Causal Inference in Understanding the Effect of Interventions

My second focus of research is applying novel methods in AI, more specifically causal inference, for clinical understanding of the decision making process from big data. The overarching goal of using data to improve human health can not be achieved, if machine learning algorithms do not untangle the mechanism of the decision making process. For example, establishing whether a causal intervention will work for a certain population is one of the fundamental challenges in the health sciences and biomedical research. The process of gathering empirical evidence to support new interventions is highly non-trivial due to the complexity of human biology and its intricate interactions with the underlying environment.

The enactment of the 1962 Amendments to the Food, Drug and Cosmetics Act required that new treatments need to be proven efficacious in "adequate and well-controlled investigations" [5]. In the 1970s, the FDA translated this into the requirement that a randomized controlled trial (RCT) is needed to validate the causal link between a new treatment (causal intervention) and a putative clinical outcome. The rationale for adopting RCTs as the means of gathering scientific evidence is that spurious associations, due to factors extraneous to the relationship between the treatment and the outcome, can be controlled for by randomizing the treatment assignment.

RCT-based procedures are currently considered the gold standard for supporting evidence generation in the empirical sciences, and dozens of trials are conducted each year just in the medical domain. Despite a host of benefits, it's widely acknowledged that RCTs are far from perfect. In practice, RCTs are usually slow and time-consuming, overly expensive, not always entirely ethical, plagued with biases, and often applicable to a narrow stratum of the population. In contrast to the RCT-based approach, scientists have been collecting increasing amounts of observational (non-experimental) data, colloquially termed *big data*. Unfortunately, the mere accumulation of high volumes of data doesn't solve all the problems either. In fact, all that can be claimed, plainly, from these big data collections are simply statistical correlations, not causation at all. The current generation of biomedical researchers face, therefore, a challenge – understand how to leverage the vast, but imperfect, amounts of data, and translate it into new insights about causal interventions.

This line of research builds on our recent work on using structure causal model to identify causation from observational data, and provide support for generalizability and transportability [3]. This research agenda takes leverage of recent advances in causal inference and graphical models [6] to address the grand challenge of inference from biased and heterogeneous data coming from various populations, task known as *data-fusion* [7]. In particular, domain knowledge (e.g., clinical understanding) is encoded in a causal graph, where nodes represent random variables (e.g., treatment, outcome, covariates) and arrows represent causal relations (including the possibility of zero-effect) [6].

One key assumption of the current framework is that experts (e.g., clinicians) have an understanding of the domain knowledge that can be translated precisely into a causal graph. This requirement may not be attainable in many settings where clinicians have only a partial understanding of the domain under investigation. While allowing the natural integration of clinical knowledge is a prominent feature of the graphical approach, requiring detailed knowledge about the causal graph may hinder its wider adoption. A critical related issue, even when the causal graph is known, is how to generalize the findings of an RCT to a population that is somewhat related to, but not the same

as the population submitted to the original intervention. If an RCT is performed in Boston, for example, a policy-maker in Los Angeles may be interested in leveraging this data to avoid the costs and harms of an RCT in LA. More challengingly, could this data be used to support inferences about the broader US population? This comes under the rubric of *external validity*, namely, how to generalize experimental findings beyond the reference population.

My goal is to provide a principled approach to evidence generation for human health improvement by taking advantage of large amounts of data using causal models. In particular, the framework provides a pathway for naturally incorporating substantive knowledge and constructing plausible causal explanations, which are two fundamental aspects in clinical decision-making. This novel methodology will lead to an accelerated process of biomedical discovery that could not be done from any specific observational or experimental (RCT) study separately.

References

[1] Mohammad Adibuzzaman, George C Kramer, Loriano Galeotti, Stephen J Merrill, David G Strauss, and Christopher G Scully. The mixing rate of the arterial blood pressure waveform markov chain is correlated with shock index during hemorrhage in anesthetized swine. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 3268–3271. IEEE, 2014.

[2] 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*, 43:137, 2016.

[3] M Bikak, M Adibuzzaman, Y Jung, Y Yih, and E Bareinboim. Regenerating evidence from landmark trials in ards using structural causal models on electronic health record. In *B104. Critical care: big data in health care- predictive analytics, clinical decision support, and rapid response*, pages A4290–A4290. American Thoracic Society, 2018.

[4] Mohammad Adibuzzaman, Poching DeLaurentis, Jennifer Hill, and Brian Benneyworth. Big data in healthcare– the promises, challenges and opportunities from a research perspective: A case study with a model database. In *AMIA Annual Symposium*, pages 384–392, 2017.

[5] Jeremy A Greene and Scott H Podolsky. Reform, regulation, and pharmaceuticalsthe kefauver–harris amendments at 50. *New England Journal of Medicine*, 367(16):1481–1483, 2012.

[6] Judea Pearl. *Causality*. Cambridge university press, 2009.

[7] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.