An Open Benchmark for Causal Inference Using the MIMIC-III Dataset

Leo Anthony Celi, M.D. Ph.D. [1], Ken Jung, Ph.D. [2], Marzyeh Ghassemi, Ph.D. [3], Carlos Guzman[4], Uri Shalit, Ph.D. [4], David Sontag, Ph.D. [4]

[1]Laboratory for Computational Physiology, Massachusetts Institute of Technology; [2]Stanford Center for Biomedical Informatics, Stanford University; [3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology; [4]Computer Science Department, Courant Institute for Mathematical Sciences, New York University

Abstract

The estimation of the causal effect of medical interventions is a core problem in clinical science. However, the lack of publicly available, curated datasets is a significant barrier to entry for causal inference researchers who wish to develop and test new methods. We address this issue by creating a curated set of healthcare causality benchmark datasets using the de-identified, publicly available MIMIC-III database [1]. This data will be publically released and advertised within the machine learning, statistics and biomedical informatics communities. Our hope is that this resource will facilitate the collaboration of the causal inference community with clinical researchers, unlocking the potential of population-wide electronic medical record systems to address important outstanding questions in clinical medicine.

Introduction

Large scale medical observational datasets hold great promise for use in inferring causal relationships between medications, comorbidities and patient outcomes [3]. For example, we might be able to infer the best anti-hypertensive treatment for a patient population for which no reliable evidence from a randomized controlled trial (RCT) exists. However, the development of statistical methods for causal inference from observational data is fraught with difficulties. Foremost among these difficulties is the problem of model validation and comparison. The only way to definitively assess the validity of a causal inference method is to compare its predictions with the outcomes of a randomized control trial.

This difficulty is an obstacle for many researchers who might otherwise be interested in working with observational data, since they cannot test their ideas with real-world data reflecting the complexity of actual clinical practice. We aim to create a common, publicly available benchmark for causal inference methods by coupling the MIMIC-III data set [1] with the results of randomized controlled trials. Specifically, we obtained a list of RCTs that have been run at the same institution as MIMIC-III over the time that the data was collected. We propose to use the results of these RCTs as estimates of ground truth causal effects. Data collected before the respective RCTs are used as observational datasets from which the respective causal effects can be modelled.

MIMIC-III is a dataset of over 58,000 hospital admissions for 38,645 adults and 7,875 neonates collected over 11 years of intensive care unit (ICU) stays at Beth Israel Deaconess Medical Center (BIDMC), a major metropolitan hospital in Boston [1]. The data for each hospital stay includes information such as demographics, vital sign measurements made at the bedside (~1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital). Unlike many EHR datasets, MIMIC-III is de-identified and publically available for academic use subject to human studies training completion. The public availability of the MIMIC-III data makes it ideal as a basis for creating a benchmark, since it would be much easier to share with the wider research community, lowering the barrier for computer scientists and statisticians who wish to work on causal inference problems.

Details

In preliminary work, we have identified all the RCTs conducted at the BIDMC ICU over the time period covered by the MIMIC-III dataset. We have evaluated these trials for feasibility of replication in MIMIC-III data. Our evaluation criteria focus on the ability to reproduce the study using the covariates available in MIMIC-III data. In particular, we evaluate the ability to replicate the inclusion/exclusion criteria, treatment assignment, and measurements of outcomes of interest.

Thus far, we have identified an initial set of 3 RCTs we believe are suitable, and we will continue our work to identify a further 3-5 for a total of 6-8 RCTs. The dataset for each of these studies will consist of outcomes, treatment assignments, outcomes, and all available covariates for the patients that precede both the treatment assignment and the initiation of the respective RCT.

For example, we identified a trial by Talmor et al. [2] that evaluates the use of esophageal pressure (PES) measurements to guide mechanical ventilation in patients with acute lung injury (ALI) or acute respiratory distress syndrome (ARDS). The inclusion criteria were patients with ALI or ARDS according to the International Consensus Conference criteria [4], excluding patients with various esophageal injuries, the treatment is the measurement of transpulmonary pressure by an esophageal balloon catheter, and the outcome of interest is the ratio of the partial pressure of arterial oxygen to the fraction of inspired oxygen at 72 hours post-ventilation. We identified 92 patients in MIMIC-III who are suitable for inclusion in a dataset for replication of the study results.

For all the studies, we will also release baseline estimates of the causal estimands using direct estimation via linear regression, along with estimates based on propensity score matching using domain experts' selections of relevant covariates. The datasets, along with the baseline models and results, will be available to researchers via a website announced to the machine learning and statistics community during the NIPS 2016 Machine Learning for Healthcare conference.

Conclusion

This work describes our in-progress efforts to create an open, public benchmark for causal inference from observational health data. Coupled with parallel efforts to transform MIMIC III into the OMOP Common Data Model, our benchmark will provide a powerful means of assessing the accuracy and trade-offs of the OHDSI methods library, such as the Population-Level Estimation tools.

References

1. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3.
2. Talmor D, Sarge T, Malhotra A, O'Donnell CR, Ritz R, Lisbon A, Novack V, Loring SH. Mechanical ventilation guided by esophageal pressure in acute lung injury. New England Journal of Medicine. 2008 Nov 13;359(20):2095.
3. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. The Cochrane Library. 2014 Jan 1.
4. Bernard GR, Artigas A, Brigham KL, Carlet J, Falke K, Hudson L, Lamy M, Legall JR, Morris A, Spragg R. The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. American journal of respiratory and critical care medicine. 1994 Mar;149(3):818-24.