Use of Electronic Health Records for Clinical Research: Issues of Study Design and Analysis

A Harvard Catalyst Biostatistics Symposium

March 1, 2018
9:00am–5:30pm

Sponsored by Harvard Catalyst’s Biostatistics Program
AGENDA

9:00am  Welcome and Opening Remarks
Rebecca Betensky, PhD
Professor of Biostatistics, Harvard T.H. Chan School of Public Health
Program Director, Harvard Catalyst Biostatistics Program

9:10am  Designing Analyses of Healthcare Databases to Emulate Randomized Trials
Miguel Hernan, MD, ScM, DrPH
Kolokotrones Professor of Biostatistics and Epidemiology, Harvard T.H. Chan School of Public Health

9:40am  Discussion

9:55am  Machine Learning on Electronic Health Records
David Sontag, PhD
Assistant Professor, Department of Electrical Engineering and Computer Science, Institute for Medical Engineering & Science, Massachusetts Institute of Technology

10:25am  Discussion

10:40am  Break

10:55am  Observational Study Designs for Highly Stratified Data
Elizabeth Mostofsky, ScD
Instructor in Epidemiology, Harvard T.H. Chan School of Public Health

11:25am  Discussion
11:40am  Adjusting for Selection Bias in Electronic Health Records-Based Research

Sebastien Haneuse, PhD
Associate Professor of Biostatistics, Harvard T.H. Chan
School of Public Health

12:10pm  Discussion

12:25pm  Lunch


Douglas MacFadden, MS
Chief Informatics Officer, Harvard Catalyst, Harvard Medical School

1:55pm  Discussion

2:10pm  Identifying Causal Effects from Electronic Health Record Data with Coarsened Exact Matching

Victor M. Castro, MS
Team Lead, Research Information Science and Computing, Partners Healthcare

2:40pm  Discussion

2:55pm  Partners eCare Research Core

Holly Barr-Vermilya, MHA
Partners eCare Research Core (PeRC) Director

Adrian Zai, MD, PhD, MPH
Partners eCare Research Director
Assistant Professor, Harvard Medical School
3:25pm  Discussion

3:40pm  Break

4:00pm  Efficient Use of EHR for Discovery Research

Tianxi Cai, ScD
Professor of Biostatistics, Harvard T.H. Chan School of
Public Health
Professor of Biomedical Informatics, Harvard Medical School

4:30pm  Discussion

4:45pm  Building Representative Matched Samples
in Large-Scale Observational Studies with
Multivalued Treatments

José R. Zubizarreta, PhD
Assistant Professor, Department of Health Care Policy,
Harvard Medical School
Faculty Affiliate, Department of Statistics, Faculty of Arts
and Sciences, Harvard University

5:15pm  Discussion

5:30pm  Closing Remarks

Rebecca Betensky, PhD
Professor of Biostatistics, Harvard T.H. Chan School of
Public Health
Program Director, Harvard Catalyst Biostatistics Program
In clinical practice, patients with the same disease diagnosis often differ in outcomes and response to treatment. The ability to both classify and predict disease phenotypes would be a valuable asset in clinical decision-making. Large datasets containing both a wealth of clinical and experimental data now exist as a result of the increasing adoption of electronic health records (EHR) linked with specimen bio-repositories. These datasets allow for data driven classification, prediction of sub-phenotypes and treatment response, as well as investigation of shared risk factors across a group of phenotypes. In this talk, I will discuss various statistical and informatics methods that illustrate both the challenges and potential opportunities that arise from analyzing EHR data.
IDENTIFYING CAUSAL EFFECTS FROM ELECTRONIC HEALTH RECORD DATA WITH COARSENED EXACT MATCHING

Victor M. Castro, MS

Team Lead, Research Information Science and Computing, Partners Healthcare

Electronic health record (EHR) data offers a real-world setting to identify new causal effects of interventions and disease progression. However, a major limitation of observational EHR data is confounding by unmeasured observations and bias associated with healthcare utilization. In this talk, we will dig into these limitations and discuss methods for identifying true causal effects from large observational EHR datasets using case-control designs with coarsened exact matching (cem). We will look at the role of computed phenotypes within this analytic framework, and we will apply the method to detect drug effects with causal links to osteoporosis by looking across 900+ medications.
ADJUSTING FOR SELECTION BIAS IN ELECTRONIC HEALTH RECORDS-BASED RESEARCH

Sebastien Haneuse, PhD

Associate Professor of Biostatistics, Harvard T.H. Chan School of Public Health

Electronic health records (EHR) data provide unique opportunities for public health and medical research. From a methodological perspective, much of the focus in the literature has been on the control of confounding bias. In contrast, selection due to incomplete data is an under-appreciated source of bias in analyzing EHR data. When framed as a missing-data problem, standard methods could be applied to control for selection bias in the EHR context. In such studies, however, the process by which data are complete for any given patient likely involves the interplay of numerous clinical decisions made by patients, healthcare providers, and the health system. In this sense, standard methods fail to capture the complexity of the data so that residual selection bias may remain.

Building on a recently-proposed framework for characterizing how data arise in EHR-based studies, sometimes referred to as the data provenance, we develop and evaluate a statistical framework for regression modeling based on inverse probability weighting that adjusts for selection bias in the complex setting of EHR-based research. We show that the resulting estimator is consistent and asymptotically normal, and derive the form of the asymptotic variance. Plug-in estimators for the latter are proposed. We use simulations to: (i) highlight the potential for bias in EHR studies when standard approaches are used to account for selection bias, and (ii) evaluate the small-sample operating characteristics of the proposed framework. Finally, the methods are illustrated using data from an on-going, multi-site EHR-based study of bariatric surgery on BMI.
Making decisions among several courses of action requires knowledge about the causal effects of each action. Randomized experiments are the preferred method to quantify those causal effects. When randomized experiments are not feasible or available, causal effects are estimated from non-experimental or observational databases. Therefore, causal inference from observational healthcare databases can be viewed as an attempt to emulate a hypothetical randomized experiment—the target experiment or target trial—that would quantify the causal effect of interest. This talk outlines a general algorithm for causal inference using healthcare databases that makes the target trial explicit. This causal framework channels counterfactual theory for comparing the effects of sustained treatment strategies, organizes analytic approaches, provides a structured process for the criticism of observational analyses, and helps avoid common methodologic pitfalls.
THE HARVARD SHRINE NETWORK: THE HARVARD CATALYST TOOL FOR QUERYING EHR DATA ACROSS HARVARD HOSPITALS

Douglas MacFadden, MS

*Chief Informatics Officer, Harvard Catalyst, Harvard Medical School*

SHRINE, the Shared Health Research Information Network, is a web-based query tool built on top of i2b2 (Informatics for Integrating Biology and the Bedside), a widely used and robust platform for clinical research. SHRINE allows researchers to query across participating hospital electronic medical record data to determine the total counts of patients who meet a given set of inclusion and exclusion criteria (currently demographics, diagnoses, medications, and selected laboratory values). The Harvard SHRINE network includes Massachusetts General Hospital, Beth Israel Deaconess Medical Center, Brigham and Women’s Hospital, Boston Children’s Hospital, and Dana-Farber Cancer Institute. Typical use cases include:

- Generating new research hypotheses
- Planning research requiring large sample sizes not easily available at any single institution
- Preparing grant applications that would benefit from pre-identification and/or characterization of a potential research cohort
- Identifying potential cohorts for clinical trials

I will also describe a few other SHRINE networks, including the national ACT network, in addition to the Harvard SHRINE network.
Electronic health datasets based on electronic health records are an increasingly popular resource for research as they provide comprehensive coverage of medication and outcome information for a large population, but information on potential confounding is often limited. Several statistical methods have been created to address confounding in large electronic health datasets commonly used for patient-centered outcomes research. These approaches use measured health characteristics to reduce confounding by between-person differences in cohort and case-control studies. Alternatively, in a case-only analysis, each individual is compared to himself/herself, eliminating confounding by stable risk factors, whether measured or not, to examine the impact of transient exposures on acute outcomes. Despite the common aim of these designs, each approach involves different assumptions, disadvantages, advantages, and analytic considerations. In this talk, I will describe the design options and the advantages of using them for analyzing electronic health records.
Electronic health records are now pervasive, presenting an incredible opportunity to use retrospective data to learn about medicine and to improve healthcare. Machine learning can help answer questions such as “What conditions does this patient have?”, “When will this patient’s disease progress?”, and “How does this treatment affect a patient’s outcomes?” However, we are often confronted with challenges such as having little labeled data, a significant amount of missing data and censoring, and the need for high-dimensional causal inference from observational data. I will discuss several new methodologies that my group has created to address these challenges, with a particular focus on disease progression modeling and estimation of individual treatment effect.
The Partners eCare Research Core (PeRC) is a core service offered by Partners eCare to support the Partners research community. It leverages the Epic EHR to assist researchers in identifying and recruiting patients for their research studies conducted at Partners HealthCare. Services that PeRC offers include: gathering Epic-related patient data for research (in the form of reports and/or data extracts), recruitment related services, and developing patient questionnaires.
In observational studies of causal effects, matching methods are widely used to approximate the ideal study that would be conducted under controlled experimentation. In this talk, I will discuss new matching methods that use tools from modern optimization to overcome four limitations of standard matching approaches. In particular, these new matching methods (i) directly obtain flexible forms of covariate balance, as specified before matching by the investigator; (ii) produce self-weighting matched samples that are representative of target populations by design; and (iii) handle multiple treatment doses without resorting to a generalization of the propensity score. These methods can handle large data sets quickly. I will illustrate the performance of these methods in an epidemiology case study about the impact of an earthquake on post-traumatic stress.
HARVARD CATALYST BIOSTATISTICS CONSULTING

Drawing on a team of highly skilled biostatisticians from Harvard and the academic healthcare centers, free consultations and expertise on a range of relevant areas are available to Harvard researchers on their clinical and translational projects. Statistical services focus on projects in the early stages of development. These services include grant submission/resubmission, IRB submission, protocol review, design for non-grant project/feasibility consultation, education on a statistical topic, analysis planning and advice, and assistance with response to a manuscript/journal reviewer. We also provide assistance with reporting results to ClinicalTrials.gov.

catalyst.harvard.edu/services/biostatsconsult/

ABOUT HARVARD CATALYST

Established in 2008, Harvard Catalyst | The Harvard Clinical and Translational Science Center is funded by the National Institutes of Health (NIH) Clinical and Translational Science Awards (CTSA) Program, grant UL1 TR001102), and is dedicated to improving human health by enabling collaboration and providing tools, training, and technologies to clinical and translational investigators. As a shared enterprise of Harvard University, Harvard Catalyst resources are made freely available to all Harvard faculty and trainees, regardless of institutional affiliation or academic degree.