Generating Real-World Evidence in Medicine (BINF 4008:004)

Logistics

Class Sessions: Tuesdays 2:10-4:00pm, PH20-200
Instructor: Tara Anand, tara.v.anand@columbia.edu
Office Hours: By appointment
TA: TBD

Course Description

Healthcare systems generate enormous volumes of data in various forms, with electronic health records and administrative claims in particular serving as rich sources of information. These real-world health data sources are potential reservoirs of insights for biomedical discovery. Real-world data, however, arises from a data-generating process, shaped by clinical workflow and billing practices, that is distinct from the true biomedical processes we aim to study. This can make distilling reliable insights from these observational data sources challenging. This course introduces the theoretical foundations and practical methods required for addressing these challenges through rigorous observational study design using real-world health data that can be used to generate reliable evidence to answer a range of clinical questions.

Each course topic is developed through three layers of theory, application in informatics, and implementation through tools developed by the Observational Health and Data Sciences and Informatics (OHDSI) consortium, which has developed analytic tools based on best practices for high-quality observational studies. The theory for this course is developed around the frameworks of probabilistic and causal inference, to facilitate understanding of study-design principles and analytic approaches that are thoughtfully informed by domain knowledge. We use theory to understand the challenges and biases of real-world health data, and how observational study design choices and methods can address them. Each stage of the observational study pipeline is covered including formulating a precise research question, identifying the right data, defining study populations and phenotypes, and determining and executing the appropriate analyses.

Tools and conventions developed by the OHDSI community will serve as a practical framework for implementing these concepts. We will examine how OHDSI tools can be used to quickly translate a well-designed research question into an executable study and to evaluate whether the resulting estimates are credible. Each class meeting includes both lecture and hands-on exercises so that concepts are translated into practice as they are introduced.

Learning Objectives

1. Understand the theoretical principles of probabilistic and causal inference
    - Distinguish between associational, interventional, and counterfactual queries
    - Construct causal diagrams representing assumptions about a data-generation process and sources of bias
    - Analyze a Bayesian network to infer independencies and dependencies
    - Evaluate causal effect identifiability from a causal diagram using key identification criteria
2. Translate theoretical principles to requirements for informatics applications with observational studies
    - Articulate potential sources of bias in EHR and claims databases
    - Formulate precise research questions with clearly specified study components
    - Define phenotypes and study cohorts to minimize measurement and other biases

- Determine appropriate analyses for a specified question

3. Utilize the OHDSI workflow including its tools as a means of implementing theory in practice.
    - Understand the OMOP common data model and standardized vocabularies
    - Construct cohort definitions in ATLAS and evaluate phenotypes
    - Execute characterization, prediction and estimation studies with Strategus
    - Evaluate results with diagnostics

Prerequisites
There are no formal course prerequisites. Students are expected to have some basic knowledge of probability and statistics and introductory coding.

Format & Grading
Weekly assignments: 25%
Final project: 50%
Participation and attendance: 25%

Weekly Assignments
Weekly assignments will primarily be completed during the in-class lab portion of each session. Class time is allocated each week for working through the assignment, which is meant to put into practice the theory and application concepts learned, and practice using OHDSI tools. If you don't complete the assignment during class time, you will complete the assignment before the next class session. Some assignments will be standardized to practice a specific skill, while other assignments will be tailored to your own final project (described below).

Final Project
You will work independently, in consultation with the instructor, to design and execute one or more OHDSI-style observational studies answering a clinically-relevant research question of your choosing. The goal is for each of you to leave this course with a manuscript draft that could be submitted for publication to a clinical journal.

You will submit two written reports of your research.

1. Journal manuscript (3,000-4,000 words): a journal-style research report with sections Introduction/Background, Methods, Results, Discussion, Conclusion. This should be formatted for a clinical journal, presenting your research question, analytical approach, results and interpretation of your findings.
2. Study design and execution reflection (1-3 pages): this reflection piece should describe your research process, including a more thorough explanation of choices made, and intermediate steps in your process that may not be reflected in your journal manuscript. This should address:
    a. What informed your phenotype/study design choices? What alternative choices could you have made, and why did you settle on the choices you did?
    b. What decisions did you make initially that may not have worked? Why did they require refinement?
    c. How did you evaluate your choices along the way?

Participation and Attendance
Attendance at all lectures is expected, with advance notice (at least 2 weeks for planned absences, and 24 hours/ASAP for illness or extenuating circumstances). Participation is

assessed through active engagement with lectures and in-class activities. Students with low-engagement will be contacted mid-way through the semester to check-in.

Policy on Generative AI

Generative AI is a new, evolving tool that we are learning how to best integrate in research. This course is an opportunity to continue this learning; you are generally encouraged to use GenAI tools as you see fit throughout the course, with some exceptions (see below). For any work you submit (weekly assignments and final project), you will be asked to share with some specificity in what ways you used GenAI.

If your use of GenAI appears to have either 1) hindered your understanding of key concepts, or 2) prevented you from submitting accurate, high-quality work, you will not be penalized, but will be given an opportunity to meet with me discuss how your use of the tool could be improved, and you will revise and resubmit the assignment with that guidance.

Policy by course component.

- **Lectures:** Slides will be posted ahead of each lecture. Recording or transcribing lectures with any tool is not permitted, except with written approval from the accommodations office.
- **Weekly assignments:** These are meant to help you 1) understand concepts and 2) practice implementing the OHDSI tool stack. You are welcome to use GenAI to help you debug code or otherwise as you see fit without hindering your comprehension.
- **Final project:** As with weekly assignments, you may use GenAI for debugging code or other support. You are encouraged to develop your ideas for your project and writeups on your own, as each has specific requirements that must be addressed with a level of accuracy and specificity that you are unlikely to meet without your own critical thinking.

Textbook / Resources

1. The Book of OHDSI
2. Causal Artificial Intelligence
3. The Book of Why

Schedule

Week #	Date	Topic
1	9/8/26	1.1 [Theory] Motivation and Introduction to observational studies in informatics 1.2 [Theory] Pearl's Causal Hierarchy 1.3 [Tool] Introduction to OHDSI
2	9/15/26	2.1 [Theory] Review of fundamentals: Probability 2.2 [Theory] Interpretation of probabilities 2.3 [Application] Domain context: EMR vs. claims data and the OMOP common data model 2.4 [Tool] ATLAS overview
3	9/22/26	3.1 [Theory] Structural causal models and Causal diagrams 3.2 [Application] Phenotyping: from diagrams to data 3.3 [Tool] ATLAS: Cohort construction part 1

4	9/29/26	4.1 [Theory] d-Separation with causal diagrams part 1- determining probabilities from diagrams 4.2 [Application] Formulating associational queries (characterization) 4.3 [Tool] Strategus Characterization package
5	10/6/26	5.1 [Theory] d-Separation with causal diagrams part 2 - reading independence structure 5.2 [Application] Formulating associational queries (prediction) 5.3 [Tool] Strategus Patient-level prediction package
6	10/13/26	6.1 [Theory] Identification part 1 - backdoor criterion and adjustment sets 6.2 [Application] Graphically encoding biases in medical data part 1 6.3 [Tool] ATLAS: Cohort construction part 2
7	10/20/26	7.1 [Theory] Identification part 2 - conditional backdoor, frontdoor criterion 7.2 [Application] Graphically encoding biases in medical data part 2 7.3 [Tool] Strategus CohortDiagnostics
8	10/27/26	8.1 [Theory] From identification to estimation - propensity-score analysis 8.2 [Application] Cohort study design 8.3 [Tool] Strategus Population-level estimation (CohortMethod) package
–	11/2/26	Election day (no class)
–	11/10/26	AMIA (no class) - Sign up for 30-min 1:1 Final Project Consult
9	11/17/26	9.1 [Application] Bias evaluation and diagnostics 9.2 [Application] Self-controlled case series study design 9.3 [Tool] Strategus SCCS package
10	11/24/26	10.1 [Theory] Transportability and data fusion 10.2 [Application] Meta-analyses and network studies 10.3 [Tool] Strategus EvidenceSynthesis Tool
11	12/1/26	Special Topics in Observational Research
12	12/8/26	Project Presentations