Welcome to BIOMEDIN 215: Data Science for Medicine

Welcome to BIOMEDIN 215! The teaching team is committed to providing a great learning experience. This syllabus will be your comprehensive roadmap to the course. If the answer to a question you have is not in the syllabus, please post on canvas or email the TAs.

Teaching Team

- Professor Nigam Shah - nigam@stanford.edu
- TAs: Joshua, Sohaib, Suhana, Yixing

Location

Lecture: Tu, Th 3-4:20 in Gates B3

Course Materials

Canvas

We will use Canvas in this course to send announcements, distribute and collect homework, and post course materials. Please ensure that you are added to BIOMEDIN 215 on Canvas. SCPD and HCP students, as well as auditors should all make sure they have access to the BIOMEDIN 215 Canvas page. If you do not have access, email the teaching team with your SUNET ID and let us know if you are an SCPD student, or an auditor. Regardless of your level of involvement with the course, you should double-check your notification settings on Canvas to ensure that you receive announcements and materials in a timely fashion. Contact the teaching team if you think you have Canvas properly set up but suspect you are not receiving notifications.

Video recordings

Video cameras located in the back of the room will capture the instructor presentations in this course. For your convenience, you can access these recordings by logging into the course Canvas site. These recordings might be reused in other Stanford courses, viewed by other Stanford students, faculty, or staff, or used for other education and research purposes. Note that while the cameras are positioned with the intention of recording only the instructor, occasionally a part of your image or voice might be incidentally captured. If you have questions, please contact a member of the teaching team.

Course Overview

The widespread adoption of electronic health records (EHRs) has created a new source of “big data”—namely, the record of routine clinical practice—as a by-product of care. Can we use this data to save lives and promote wellbeing?

Learning Goals

Upon completing this course, you should be able to:

1. differentiate between and give examples from categories of research questions and the study designs used to address them.
2. describe common healthcare data sources and their advantages and limitations relative to different research questions.
3. extract and transform various kinds of clinical data to create analysis-ready datasets.
4. select and implement a statistical analysis that is appropriate for a given dataset and research question.
5. apply your judgment to assess the accuracy and value of clinical informatics research.

The overall goal of this course is to prepare you to discover meaningful clinical knowledge using healthcare data. In addition, the practical skills you will learn in this class will be applicable to any task involving healthcare data manipulation and analysis.

What we expect from you

We are confident that every one of you will succeed and depart with a greatly expanded understanding of clinical data analysis. Our standards in the course are high, but that is because we know from experience that you can meet them. We are here to guide you in your growth. We recognize that mistakes are the core of learning, so we encourage you to make them!

Since we are confident in your integrity we also hold each of you to the highest ethical standards. We hope to motivate you to be the best person and scientist you can be. We expect you to follow the honor code and fundamental standard at all times.

Prerequisites

Must-Have: This course is designed for students experienced in basic programming, statistics, and biology. To ensure your success in this course, we expect that you know the R programming language and require that you have taken the following courses:

- CS 106A or equivalent
- STATS 60 or equivalent
- high school biology

Helpful: In this course you will use concepts from machine learning, data management, biostatistics, and physiology to answer clinical questions. While very useful, prior knowledge of these topics is not strictly required because this class focuses on the proper application and contextualization of these concepts. There are many outstanding classes at Stanford that provide a deeper dive into these topics, some of which are listed here:

- CS 106B
- STATS 216, STATS 315A, or CS 229
- CS 145 or CS 246
- STATS 305A or HRP 258
- BIO 112

Getting up to speed in R: We will use the R programming language to perform data transformations, analyses, and visualization. In particular, we will focus on the immensely useful tidyverse packages. If you've used base R before, you'll be amazed at how tedious tasks can be boiled down to a few lines of easily readable dplyr, purrr, or tidyr calls. The definitive resource for data-munging with R is *R for Data Science*. We recommend you read through this book in the first few weeks of class, since it will come in handy for the homework. We will use the caret package to provide a consistent interface to the predictive models we will study in the last part of the quarter.

Statistical Inference: We will review some statistical theory in class, but our focus in this class is how to correctly interpret results. To deepen your knowledge, we recommend working through Stanford's excellent online course: Introduction to Probability and Statistics for Epidemiology. You can sign up directly with your SUNET ID. It is not necessary to watch all of the lectures: we recommend taking the quizzes for each section and starting the course at whatever point your knowledge starts to fall off.

Predictive Modeling: As with statistical inference, our focus in teaching predictive models is on how to use them and how to interpret results. An introduction to statistical learning provides a friendly introduction to many machine learning models and their inner workings.

Lectures

Lectures will present the course material in an engaging, interactive (and hopefully fun) way. If you do not attend lectures, you will find it more difficult to internalize key concepts and do well on the homework assignments.

The lecture schedule is at https://shahlab.stanford.edu/biomedin215.

Deliverables and Grading

Homework (60%): The purpose of the homework is for you to practice doing your own clinical informatics studies. Homework will be assigned and due online through Canvas. Assignments count for a substantial portion of your grade because doing them is when you will be learning the most. Homeworks 1, 2, 5 and 6 are worth 10% of the course grade. Homework 3 is worth 8% of the course grade, and homework 4 is worth 12% of the course grade. We recommend you start early and do not leave all the work to the last minute.

You are encouraged to collaborate in pairs on the homework because that is more productive for learning. The best way to collaborate is to work together through each problem - the way the assignments are structured makes it difficult to split up the problems (and your learning will be less effective if you do that). Every person must turn in their own write-up, i.e. each student should write and run their own code on their own machine. Clearly identify who you worked with on your write-up.

All assignments will be distributed as markdown .md and .ipynb documents, which you can edit directly in a markdown editor, jupyter notebook, or in Rmarkdown/knitr through Rstudio. This is a great way to practice using the tools of reproducible research. In fact, all of the assignments and this syllabus itself were developed and written in jupyter notebooks! Please submit all homework through Canvas as .pdf files. All of the above tools allow for an easy export to .pdf, but there are bugs now and again. Please test and debug to make sure you are able to properly export your files before the due date of the assignment.

We will often grade liberally with partial credit, so we encourage you to try every problem. Even writing the logic of what you would do to solve the problem in plain english is helpful! Solution sets will be available during TA office hours for you to examine and ask questions about - we highly encourage that you visit office hours to get more feedback on your graded assignments.

Late Policy: Each of you is allowed three late days for the whole quarter to use for any homework assignment or combination of homework assignments. If you use a late day, you will not be penalized any points. If you do not use late days, each additional day of tardiness will cost you 20% of the total possible points on the assignment. Unused late days will be counted towards participation credit.

MIMIC Data Use Agreement: Assignments 3, 4, 5, 6 will require you to use a publicly available dataset called MIMIC. To access it, you are legally required to sign a data use agreement (DUA). The process is straightforward, but it may take a few days, so you should do this ASAP. More details can be found on the MIMIC DUA assignment sheet on Canvas.

Take Home Paper Review (20%): Your final opportunity to practice being a clinical informatics researcher in this course will be to write a review of a paper that we will give you. You will take the place of a reviewer to whom the paper has been sent by a journal editor. It will be your job to decide if the paper should be accepted, reviewed (major or minor), or rejected, and on what grounds.

In Class Exercise on Phenotyping and Cohort Building (10%): This hands-on session will get you comfortable using Stanford Medicine computational resources including the Nero secure compute environment and Jupyter notebooks, as well as defining phenotypes and building cohorts using the OHDSI ATLAS tool. You will conduct short coding exercises to query, summarize and analyze patient-level data from the STARR OMOP database, and discuss approaches and challenges.

Participation (10%): Participation is graded to encourage you to be active learners who engage with the curriculum. Half of your participation grade (5%) comes from meaningfully filling out the post-lecture surveys on Canvas. These are available after each lecture and you can complete them within 3 days after the lecture. In order to accommodate diverse schedules and needs the second half of the participation grade is flexible - you can earn full points by doing any of the following:

- Asking questions during the class
- Attending class regularly (18 out of 20 counts as 'full' attendance)
- Making a meaningful contribution on Canvas Forums related to topics discussed in class

Office Hours. Office hours are an opportunity for you to connect with TAs and get your questions answered. We encourage you to be engaged in your learning and take advantage of office hours - the TAs are friendly and inclusive and are happy to help out or point you in the right direction. The TAs will hold joint office hours as posted on the class Canvas site. Other times may be available by appointment to accommodate for special circumstances and class schedules.

Readings. Readings will reinforce the material from lectures, provide further details, and give you insight into how researchers in the field think about key concepts. We will have occasional assigned readings, which will be announced in class as the last lecture slide (if there is assigned reading). The relevant pdf will be available under files in Canvas.

Canvas. All course communication and material will come through Canvas. Canvas Forums are a great way to interact with your fellow students to get your questions answered and form collaborations.

An Inclusive and Equitable Learning Environment

We consider your diversity to be a strength and a positive contribution to this learning community. We expect every member of our community to contribute to an inclusive and respectful culture. Dimensions of diversity can include sex, race, age, national origin, ethnicity, gender identity and expression, intellectual and physical ability, sexual orientation, income, faith and non-faith perspectives, socio-economic class, political ideology, education, primary language, family status, military experience, cognitive style, and communication style. The individual intersection of these experiences and characteristics must be valued in our community. We will use inclusion strategies

inside and outside the classroom to make sure that everyone's contributions are welcomed and valued.

Students with Documented Disabilities: Students who may need academic accommodation based on the impact of a disability must initiate the request with the Office of Accessible Education (OAE). Professional staff will evaluate the request with required documentation, recommend reasonable accommodations, and prepare an Accommodation Letter for faculty dated in the current quarter in which the request is being made. Students should contact the OAE as soon as possible since timely notice is needed to coordinate accommodations. The OAE is located at 563 Salvatierra Walk (phone: 723-1066, URL: http://oae.stanford.edu).

Computing Resources: We have designed all homework to be runnable on a personal laptop. If you do not have regular access to a reasonably fast personal computer, you can use your own machine or library computers to access compute resources through the Stanford Shared Computing resources.

Course Feedback

We strongly believe that engaging fully in the class will help you reach the learning goals, but we continue to refine our processes based on student feedback; and there may be bumps along the way. We appreciate your ongoing feedback and patience with any problems or inconsistencies throughout the course. Your reflection and input will strengthen your learning and improve the class in years to come. You will have an opportunity to provide structured feedback throughout the course and via a review survey at the end of the quarter.

HCP and NDO Students

HCP and NDO students can fully participate in every aspect of BIOMEDIN 215. Students can watch live lectures through Panopto Course Videos or come in person. Additionally students can contribute to canvas discussions. Students should submit all of their homework assignments exclusively through Canvas.