Public Health 290 — Spring 2018 Syllabus
Targeted Learning in Biomedical Big Data

Class meets TuTh 11:00A–12:30P in Mulford 230
Lab meets W 2:00–3:00P in Mulford 230

Section 011 | Course Control Number: 42472

Instructor: Mark van der Laan
E-mail: laan@berkeley.edu
Web: https://vanderlaan-group.github.io
Office Location: 108 Haviland Hall
Office Hours: Tu 4:00–5:00P

GSI: Nima Hejazi
E-mail: nhejazi@berkeley.edu
Web: https://nimahejazi.org
Office Location: 111 Haviland Hall
Office Hours: M 2:00–3:00P

The instructional staff reserve the right to make changes to the syllabus at any time.

Course Description: This course teaches students how to construct efficient estimators and obtain robust inference for parameters that utilize data-adaptive estimation strategies (i.e., machine learning). Students will perform hands-on implementation of novel estimators using high-dimensional biomedical data sets, providing students with a toolbox for analyzing complex longitudinal, observational, and randomized control trial data. Students will actively learn and apply the core principles of the Targeted Learning methodology, which (1) generalizes machine learning to any estimand of interest; (2) obtains an optimal estimator of the given estimand, grounded in theory; (3) integrates the state-of-the-art ensemble machine learning techniques; and (4) provides formal statistical inference in terms of confidence intervals and testing of specified null hypotheses of interest. It also integrates causal inference thereby allowing one to define estimands that represent the answer to causal questions of interest.

Instructional Strategy: Most pedagogical studies (i.e., those concerned with the methods and effectiveness of teaching) indicate that lectures by themselves are a poor way of engaging students and promoting learning. To address this problem, this course will use a Blended Learning/Hybrid Classroom format. This involves shifting the majority of the material presented in class to out of class. Instructional core content is delivered online, outside of the classroom. Class time is spent exploring topics in greater depth and creates meaningful learning opportunities. This rearrangement allows for more interactive, active learning opportunities during class time like group discussion, Q&A, problem solving activities, and labs where students will apply the methods presented to real data. It also allows for self-paced comprehension of highly complex core concepts. Video lectures give students the ability to pause, rewind, and even re-watch content delivery opposed to traditional lectures that require content delivery to occur in a fixed time and place. The video lectures are intended to serve as jumping off points to drive discussion, activities, and clarification during class time. Thus, it is essential that students watch the video lecture and take notes *before* class. Under this instructional model, coming to class confused is welcome, but coming to class empty-headed is ineffective.

Prerequisites:

- Public Health C240A / Statistics C245A.
- Statistics 201A-B (or older version Statistics 200A-B).
- *recommended:* Statistics C239A / Political Science C236A or Public Health 252D or equivalent introduction to statistical causal inference.

Credit Hours: 4

Requirements and Materials:

There will be assigned and optional readings from the following textbooks that serve as excellent references, encompassing all of the topics covered in this course and more:

- *Targeted Learning: Causal Inference for Observational and Experimental Data* by Mark J. van der Laan and Sherri Rose (2011) {abbreviated to vdL&R (2011) in the sequel}
- *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies* by Mark J. van der Laan and Sherri Rose (2017) {abbreviated to vdL&R (2017) in the sequel}

These textbooks are freely available in PDF for students to download through SpringerLink.

We encourage the use of Mark van der Laan's blog (`https://vanderlaan-group.github.io/post/`) for web-based discussion. Students are invited to submit a question to the blog by sending an email to `vanderlaan.blog@berkeley.edu` at any time. Course assignments will require submission of questions to the blog.

Lecture videos, lab activity materials, homework assignments, upcoming due dates, the syllabus, and student grades will be made available via bCourses or a public website. There will be an anonymous mid-semester feedback survey; the link will be sent to students' `berkeley.edu` email addresses. This survey will be provided through SurveyMonkey or a similar service.

Learning Outcomes:

At the completion of this course, students should be able to

1. utilize the R statistical software to do the following to either longitudinal, observational, or randomized trial biomedical data:
    - apply the Super Learner algorithm to estimate various functional target parameters,
    - implement Targeted Maximum Likelihood Estimation to construct an efficient estimator of a target parameter of interest and obtain statistical inference via confidence intervals,
    - design data simulations to practically evaluate these estimators and their inference;
2. formulate the statistical estimation problem using formal notation by defining scientific question in terms of a structural causal model, the type of intervention, and the causal parameter of interest;
3. establish identifiability of the causal parameter from the observed data distribution;
4. understand the framework of the Super Learner approach for the estimation of parameters and oracle inequalities for the general cross-validation selector;
5. evaluate the conditions guaranteeing asymptotic linearity and efficiency of the TMLE for a diverse set of commonly encountered, real-world estimators of interest.

Laboratory Section:
The weekly laboratory section is intended to provide practical experience in applying the methodology discussed in the lecture videos and elaborated on in class meetings. Consistent with course prerequisites, students are expected to have advanced training in probability and statistics, to have a working knowledge of core topics in statistical causal inference, to be reasonably fluent in R, and to have some familiarity with version control and unit testing. The lab consists primarily of applying the robust statistical procedures we discuss, rather than *talking about* how to apply these statistical methods. Students are expected to have the motivation and intellectual maturity to acquire any technical skills and tools they might lack to succeed in the course (without significant reliance on the instructor), as that is what is required of professionals. We will introduce a selection of the topics mentioned above; review and elaborate on material from the lecture component; and, most importantly, address how to make use of the Targeted Learning framework to answer scientific questions in a manner emphasizing both transparency and computational reproducibility.

Grading:

- *Assignments (**40%**):* Concern the application of the above learning outcomes. Assignments incorporate written problems and programming in R. Assignments will be distributed via GitHub Classroom, and no late assignments will be accepted. There will be between three and five homework assignments.
- *Final Project (**30%**):* Consists of the presentation of a topic that involves the application of statistical methods and software to address a particular question of interest.
- *Participation (**20%**):* Includes engagement during class, coming prepared to class, team-based learning exercises, a blog question assignment, and course evaluations. Note that preparedness for class will be assessed in the form of short, random concept checks that are open note and easy for students who listened to the lecture video before class.
- *Attendance (**10%**):* Recorded at all lab and lecture meetings. You must email Prof. van der Laan, not the GSI, if you cannot attend class.

Course Policies:

Accommodations: Please speak with the instructional staff as soon as possible if you require any particular accommodations, and we will work out the necessary arrangements.

Scheduling Conflicts: Notify the instructional staff by the second week of the term about any known or potential conflicts (e.g., religious observances, interviews, team activities, conferences).

Collaboration and Independence:
All homework assignments should clearly list collaborators and references. Homework assignments will not be considered for credit if they are a replicate of another classmate's assignment. With that in mind, you may work together but you should complete answers independently.

Honor Code:

"*As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others.*"

The purpose of the Honor Code is to enhance awareness of the need for the highest possible levels of integrity and respect on campus, both within and outside the academic context. We hope and believe that the code will catalyze a series of ongoing conversations about our principles and practices. Together, through engagement, we can create a consistent message and ethos in our classrooms, labs, departments, and throughout the academic enterprise, to ensure that the core values of academic integrity and honesty are being embraced by both students and faculty. Please carefully read the Honor Code (`http://asuc.org/honorcode/index.php`).

Academic Integrity:
One of the most important values of an academic community is the balance between the free flow of ideas and the respect for the intellectual property of others. Researchers don't use one another's research without permission; scholars and students always use proper citations in papers; professors may not circulate or publish student papers without the writer's permission; and students may not circulate or post materials (handouts, exams, syllabi — any class materials) from their classes without the written permission of the instructor.
Any test, paper or report submitted by you and that bears your name is presumed to be your own original work that has not previously been submitted for credit in another course unless you obtain prior written approval to do so from yourinstructor. In all of your assignments, you may use words or ideas written by other individuals in publications, web sites, or other sources, but only with proper attribution. If you are not clear about the expectations for completing an assignment or taking a test or examination, be sure to seek clarification from your instructor or GSI beforehand. Finally, you should keep in mind that as a member of the campus community, you are expected to demonstrate integrity in all of your academic endeavors and will be evaluated on your own merits. The consequences of cheating and academic dishonestyincluding a formal discipline file, possible loss of future internship, scholarship, or employment opportunities, and denial of admission to graduate school are simply not worth it.

Students with disabilities: If you require accommodations, please make arrangements in a timely manner through the DSP office.

Important Due Dates:

Homework 1	Thursday, Feb. 15
Homework 2	Thursday, Mar. 01
Homework 3	Thursday, Mar. 15
Final Project Proposal	Thursday, Mar. 22
Mid-semester Feedback Survey	Friday, Mar. 23
Homework 4	Thursday, Apr. 05
Homework 5	Thursday, Apr. 19
Blog Question	Friday, Apr. 20
Final Project Presentations	Thursday, May 03
Final Project Report	Friday, May 04

Tentative Course Outline

The weekly coverage is subject to change, as it depends on the progress of the class.

Week	Content
1: 16–18 Jan.	• *Topics:* The roadmap of statistical learning • *Before:* Read course syllabus; watch Week 1 video lectures • *Lab:* Reproducible Research with R, git, and GitHub (part 1)
2: 23–25 Jan.	• *Topics:* Examples of data-generating experiments, traditional data analysis • *Before:* Read Ch. 1 of vdL&R (2011); watch Week 2 video lectures • *Lab:* Reproducible Research with R, git, and GitHub (part 2)
3: 30 Jan.–01 Feb.	• *Topics:* Structural causal models, causal quantities, identification • *Before:* Read Ch. 2 of vdL&R (2011); watch Week 3 video lectures • *Lab:* Structural causal models, interventions and identifiability (part 1)
4: 06–08 Feb.	• *Topics:* Interventions, optimal interventions, and identifiability results • *Before:* Watch Week 4 video lectures • *Lab:* Structural causal models, interventions and identifiability (part 2)
5: 13–15 Feb.	• *Topics:* Understanding the challenges of nonparametric density estimation: Super Learning of a density • *Before:* Read Ch. 3 of vdL&R (2011); watch Week 5 video lectures • *Lab:* Introduction to Super Learning and the `sl3` R package • *Deliverables:* Homework assignment 1 due 15 Feb. by 11:59pm
6: 20–22 Feb.	• *Topics:* Super Learner and an oracle inequality for the general cross-validation selector, Super Learning in prediction, Super Learning of optimal individualized treatment rules • *Before:* Watch Week 6 video lectures • *Lab:* Super Learner libraries and screening algorithms
7: 27 Feb.–01 Mar.	• *Topics:* Super Learning of conditional multinomial distributions or densities • *Before:* Read Ch. 15 of vdL&R (2017); watch Week 7 video lectures • *Lab:* Super Learning of densities with the `sl3` and `condensier` R packages • *Deliverables:* Homework assignment 2 due 01 Mar. by 11:59pm
8: 06–08 Mar.	• *Topics:* Measure-theoretic Integration, the Highly Adaptive Lasso (HAL) • *Before:* Read Ch. 6 of vdL&R (2017); watch Week 8 video lectures • *Lab:* The Highly Adaptive Lasso and the `hal9001` R package
9: 13–15 Mar.	• *Topics:* Asymptotic linearity, influence curves, and statistical inference based on influence curves • *Before:* Read A.1–A.2 of vdL&R (2011); watch Week 9 video lectures • *Lab:* Estimation and inference with influence functions (part 1) • *Deliverables:* Homework assignment 3 due 15 Mar. by 11:59pm

Tentative Course Outline (continued...)

Week	Content
10: 20–22 Mar.	• *Topics:* Pathwise differentiable target parameters, gradients and the canonical gradient of infinite-dimensional models • *Before:* Read A.3 of vdL&R (2011); watch Week 10 video lectures • *Lab:* Estimation and inference with influence functions (part 2) • *Deliverables:* Final project proposals due 22 Mar. by 11:59pm; Mid-semester feedback survey due 23 Mar. by 11:59pm
27–29 Mar.	• *Topics:* None — it's Spring Break • *Before:* No reading • *Lab:* Canceled
11: 03–05 Apr.	• *Topics:* Definition of MLEs and NP-MLEs, efficient influence curves, theorems of efficiency • *Before:* Read A.4 of vdL&R (2011); watch Week 11 video lectures • *Lab:* Super Learning for surival analysis • *Deliverables:* Homework assignment 4 due 05 Apr. by 11:59pm
12: 10–12 Apr.	• *Topics:* Efficient one-step estimators, online one-step estimators • *Before:* Read Ch.4–6 and A.6 of vdL&R (2011); watch Week 12 video lectures • *Lab:* Computing Targeted Maximum Likelihood Estimators and the `tmle3` R package
13: 17–19 Apr.	• *Topics:* Targeted Maximum Likelihood Estimation (TMLE) • *Before:* Read Ch. 3–4 of vdL&R (2017); watch Week 13 video lectures • *Lab:* Targeted Learning for surival analysis with the `survtmle` R package • *Deliverables:* Homework assignment 5 due 19 Apr. by 11:59pm; Blog question due 20 Apr. by 11:59pm
14: 24–26 Apr.	• *Topics:* TMLEs of causal effects of multiple time-point interventions based on longitudinal data • *Before:* Watch Week 14 video lectures • *Lab:* Special Topics: Variable Importance, CV-TMLE, TMLEs with Doubly Robust Inference
15: 01–03 May	• *Topics:* RRR week — Final project presentations • *Deliverables:* Final Project Presentations on 03 May, time TBA; Final Project Reports due 04 May by 5:00pm