To: Organizing Committee, Atlantic Causal Inference Conference 2019

From: Prof. Mark J. van der Laan

Subject: Workshop proposal: The `tlverse` software ecosystem for causal inference.

Date: December 11, 2018

---

1 Workshop information

1.1 Title: The `tlverse`: A Software Ecosystem for Causal Inference and Targeted Learning

1.2 Abstract

This full-day workshop will provide a comprehensive introduction to both the `tlverse` software ecosystem and the field of targeted learning for causal inference. While this will primarily be a software workshop centered around the new `tlverse` ecosystem (`https://github.com/tlverse`) of `R` packages, there will be rigorous examination of both causal inference methodology — focusing on the field of targeted learning — and applications in both large-scale observational studies and randomized experiments. The focus will be on introducing modern methodological developments in statistical causal inference and their corresponding software implementations in the `tlverse`. Through vignette-guided live coding exercises, participants will perform hands-on implementation of novel estimators for assessing causal claims with complex, observational data. Topics to be addressed include ensemble machine learning; efficient substitution estimators in nonparametric and semiparametric models and targeted minimum loss-based estimation (TMLE); inference based on influence functions; static, dynamic, optimal dynamic, and stochastic interventions. TMLE provides a strategy for constructing (double) robust efficient plug-in estimators with normal limiting distributions, allowing for valid inference even when the functional nuisance parameters are estimated via machine learning. Causal parameters and corresponding estimators will be examined both mathematically and through their corresponding `R` package implementations from the `tlverse` ecosystem via hands-on data analysis, providing participants opportunities to develop skills that will translate to real-world causal inference analyses. Some background in mathematical statistics will be useful; familiarity with the `R` programming language will be essential.

1.3 Motivation

Randomized clinical trials (RCTs) have long served as the gold standard of evidence for comparing potential interventions in clinical medicine and public health, marketing, political science, and a great many other fields. Unfortunately, such trials are often not feasible due to ethical, logistic or economical constraints. Observational studies constitute a potentially rich alternative to RCTs, providing an opportunity to learn about the causal effects of interventions for which little or no trial data can be produced; however, in such studies intervention allocation may be strongly confounded by other important characteristics. Thus, great care is needed in attempts to disentangle observed relationships and, ultimately, infer causal effects. This workshop will provide a comprehensive introduction to the field of targeted learning, a modern statistical framework that utilizes state-of-the-art machine learning to flexibly adjust for confounding while yielding efficient, unbiased estimators and valid statistical inference, thus unlocking observational studies for causal inference.

Targeted learning is a complex statistical approach and, in order for this method to be accessible in practice, it is crucial that it is accompanied by robust software. The `tlverse` software ecosystem was developed to fulfill this need. Not only does this software facilitate computationally reproducible and efficient analyses, but it is also a tool for targeted learning education since its workflow mirrors that of the methodology. That is, the `tlverse` paradigm does not focus on implementing a specific estimator or a small set of related estimators — instead, the focus is on exposing the statistical framework of targeted learning itself! Thus, users are required to explicitly define objects to model key statistical objects: the nonparametric structural equation model, the factorized likelihood, counterfactual interventions, causal parameters, and algorithmic step for computing estimators. All `R` packages in the `tlverse` ecosystem directly model the key objects defined in the mathematical and theoretical framework of targeted learning. What's more, the `tlverse` `R` packages share a core set of design principles centered on extensibility, allowing for them to be used in conjunction with each other and built upon one other in a cohesive fashion.

1.4 Duration

This will be a full-day workshop, featuring modules that introduce a distinct causal question motivated by a case study, alongside statistical methodology and software for assessing the causal claim of interest. A sample schedule would take the form:

- 09:30AM–10:15AM: Introduction to targeted learning for causal inference
- 10:15AM–10:45AM: Introduction to the `tlverse` software ecosystem
- 10:45AM–11:00AM: Coffee Break
- 11:00AM–11:45AM: Ensemble machine learning with the `sl3` package
- 11:45AM–12:30PM: Targeted learning for causal inference with the `tmle3` package
- 12:30PM–1:30PM: Lunch
- 01:30PM–02:45PM: Optimal treatments regimes and the `tmle3mopttx` package
- 02:45PM–03:00PM: Coffee Break
- 3:00PM–4:00PM: Stochastic interventions and the `tmle3shift` package
- 04:00PM–4:30PM: Course summary and concluding remarks

Please note that the workshop will be 6 hours, including coffee breaks but not lunch.

1.5 Prior History

The `tlverse` ecosystem is a relatively recent effort, about 2 years in the making. Although some material has been introduced across several graduate-level courses taught at UC Berkeley, this workshop would be the first offering in the 6-hour format.

2 Organizers

Mark van der Laan, Ph.D.

Mark van der Laan, PhD, is Professor of Biostatistics and Statistics at UC Berkeley. His research interests include statistical methods in computational biology, survival analysis, censored data, adaptive designs, targeted maximum likelihood estimation, causal inference, data-adaptive loss-based learning, and multiple testing. His research group developed loss-based super learning in semiparametric models, based on cross-validation, as a generic optimal tool for the estimation of infinite-dimensional parameters, such as nonparametric density estimation

and prediction with both censored and uncensored data. Building on this work, his research group developed targeted maximum likelihood estimation for a target parameter of the data-generating distribution in arbitrary semiparametric and nonparametric models, as a generic optimal methodology for statistical and causal inference. Most recently, Mark's group has focused in part on the development of a centralized, principled set of software tools for targeted learning, the `tlverse`. Contact: `laan@berkeley.edu`.

Alan Hubbard, Ph.D.

Alan Hubbard is Professor of Biostatistics, former head of the Division of Biostatistics at UC Berkeley, and head of data analytics core at UC Berkeley's SuperFund research program. His current research interests include causal inference, variable importance analysis, statistical machine learning, estimation of and inference for data-adaptive statistical target parameters, and targeted minimum loss-based estimation. Research in his group is generally motivated by applications to problems in computational biology, epidemiology, and precision medicine. Contact: `hubbard@berkeley.edu`.

Jeremy Coyle, Ph.D.

Jeremy Coyle is a consulting data scientist and statistical programmer, currently leading the software development effort that has produced the `tlverse` ecosystem of `R` packages and related software tools. Jeremy earned his PhD in Biostatistics from UC Berkeley in 2016, primarily under the supervision of Alan Hubbard. Contact: `jeremyrcoyle@gmail.com`.

Nima Hejazi, M.A.

Nima is a PhD candidate in biostatistics with a designated emphasis in computational and genomic biology, working jointly with Mark van der Laan and Alan Hubbard. Nima is affiliated with UC Berkeley's Center for Computational Biology and NIH Biomedical Big Data training program. His research interests span causal inference, nonparametric inference and machine learning, targeted loss-based estimation, survival analysis, statistical computing, reproducible research, and high-dimensional biology. He is also passionate about software development for applied statistics, including software design, automated testing, and reproducible coding practices. Contact: `nhejazi@berkeley.edu`.

Ivana Malenica, M.A.

Ivana is a Ph.D. student in biostatistics advised by Mark van der Laan. Ivana is currently a fellow at the Berkeley Institute for Data Science, after serving as a NIH Biomedical Big Data and Freeport-McMoRan Genomic Engine fellow. She earned her Master's in Biostatistics and Bachelor's in Mathematics, and spent some time at the Translational Genomics Research Institute. Very broadly, her research interests span non/semi-parametric theory, probability theory, machine learning, causal inference and high-dimensional statistics. Most of her current work involves complex dependent settings (dependence through time and network) and adaptive sequential designs. Contact: `imalenica@berkeley.edu`.

Rachael Phillips, M.A.

Rachael is a Ph.D. student in biostatistics, advised by Alan Hubbard and Mark van der Laan. She has an M.A. in Biostatistics, B.S. in Biology with a Chemistry minor and a B.A. in Mathematics with a Spanish minor. Rachael is motivated to solve real-world, high-dimensional problems in human health. Her research interests span causal inference, machine learning, nonparametric statistical estimation, and finite sample inference. She is also passionate about online mediated education. Rachael is affiliated with UC Berkeley's Center for Computational Biology, NIH Biomedical Big Data Training Program, and Superfund Research Program. Contact: `rachaelvphillips@berkeley.edu`.