Targeted Learning: Causal Inference Meets Ensemble Machine Learning

Alan Hubbard[1] Mark van der Laan[1]

[1]Division of Biostatistics, University of California at Berkeley

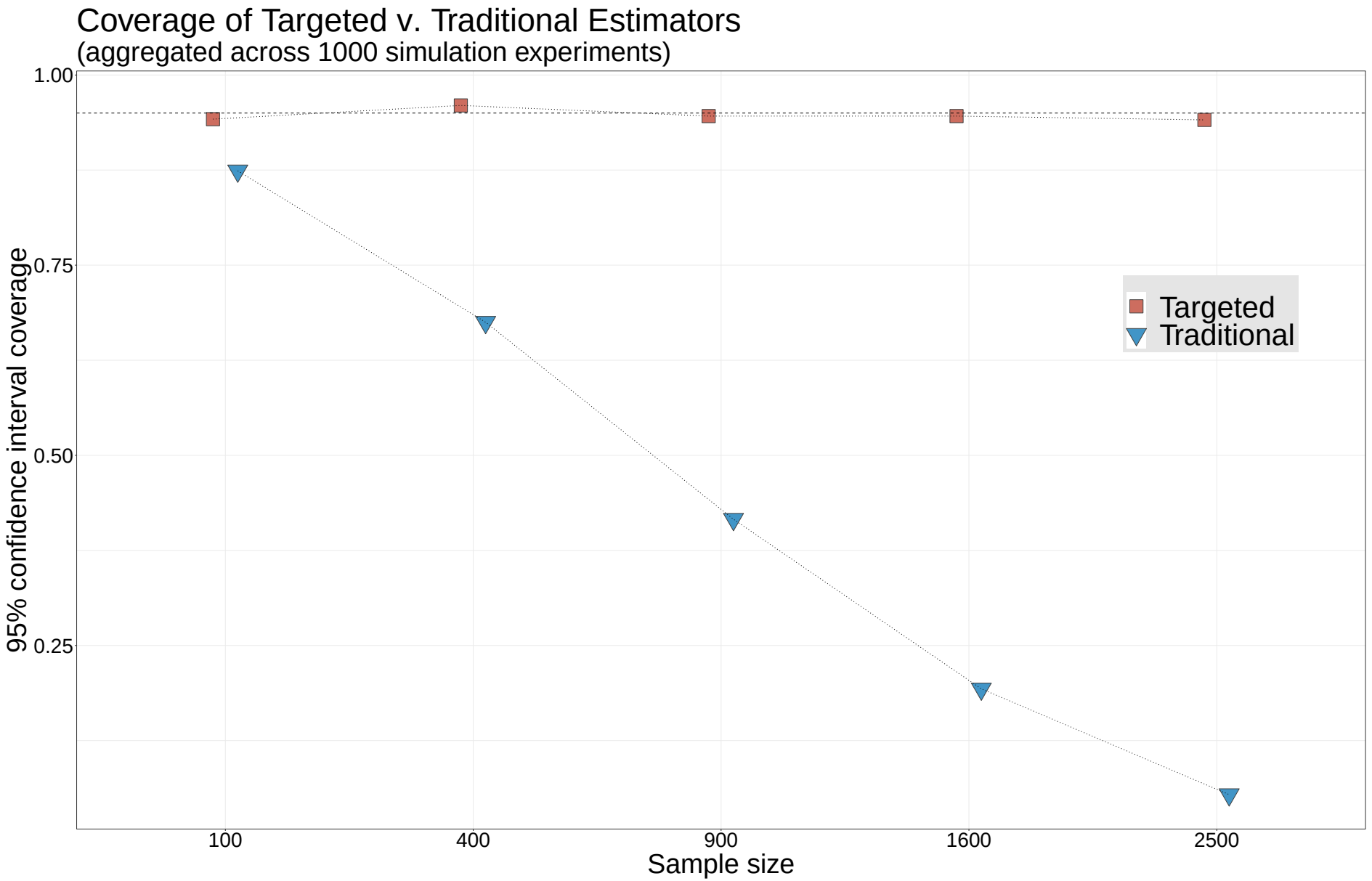
August 16, 2021
Public Health and Epidemiology Program (PASPE) at the National Institute of Public Health in Mexico

Outline

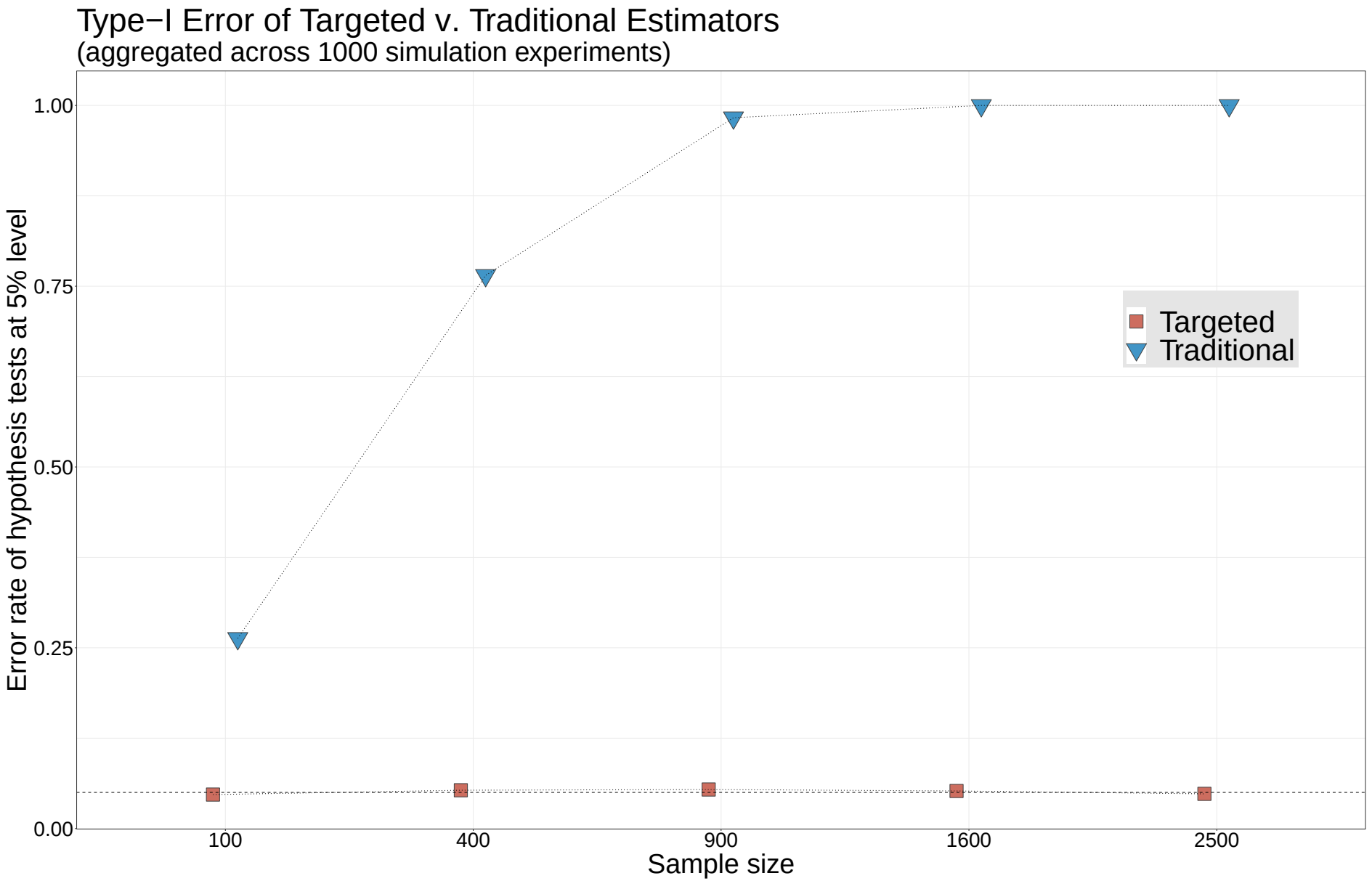
Traditional toolbox for statistics

	Type of Data			
Goal	**Measurement (from Gaussian Population)**	**Rank, Score, or Measurement (from Non- Gaussian Population)**	**Binomial (Two Possible Outcomes)**	**Survival Time**
Describe one group	Mean, SD	Median, interquartile range	Proportion	Kaplan Meier survival curve
Compare one group to a hypothetical value	One-sample *t*test	Wilcoxon test	Chi-square or Binomial test **	
Compare two unpaired groups	Unpaired *t* test	Mann-Whitney test	Fisher's test (chi-square for large samples)	Log-rank test or Mantel-Haenszel*
Compare two paired groups	Paired *t* test	Wilcoxon test	McNemar's test	Conditional proportional hazards regression*
Compare three or more unmatched groups	One-way ANOVA	Kruskal-Wallis test	Chi-square test	Cox proportional hazard regression**
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test	Cochrane Q**	Conditional proportional hazards regression**
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients**	
Predict value from another measured variable	Simple linear regression or Nonlinear regression	Nonparametric regression**	Simple logistic regression*	Cox proportional hazard regression*
Predict value from several measured or binomial variables	Multiple linear regression* or Multiple nonlinear regression**		Multiple logistic regression*	Cox proportional hazard regression*

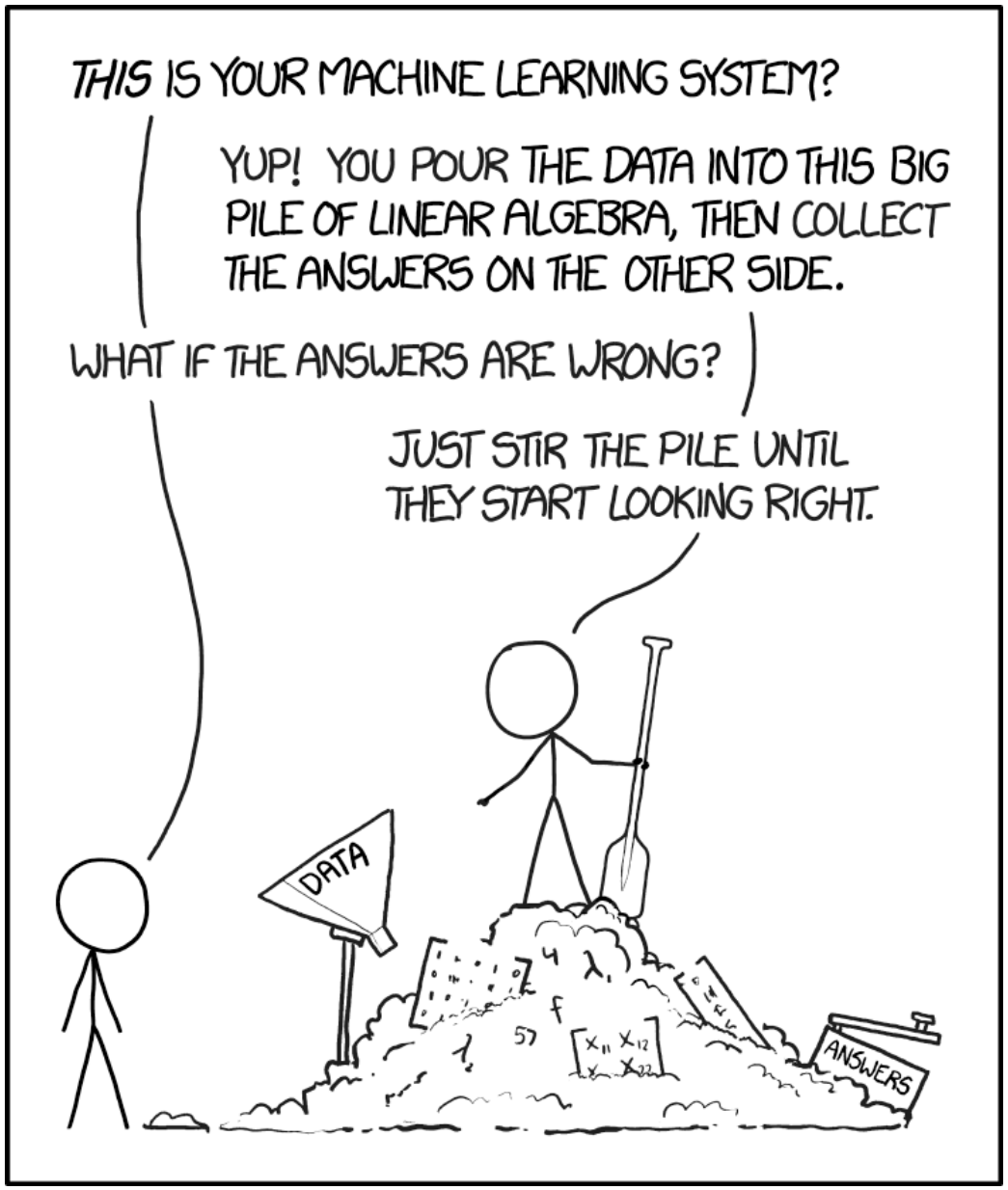
Performance of traditional tools



Performance of traditional tools



Post-hoc model manipulation



Why care about statistical inference?

Why Most Published Research Findings Are False

John P. A. Ioannidis

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

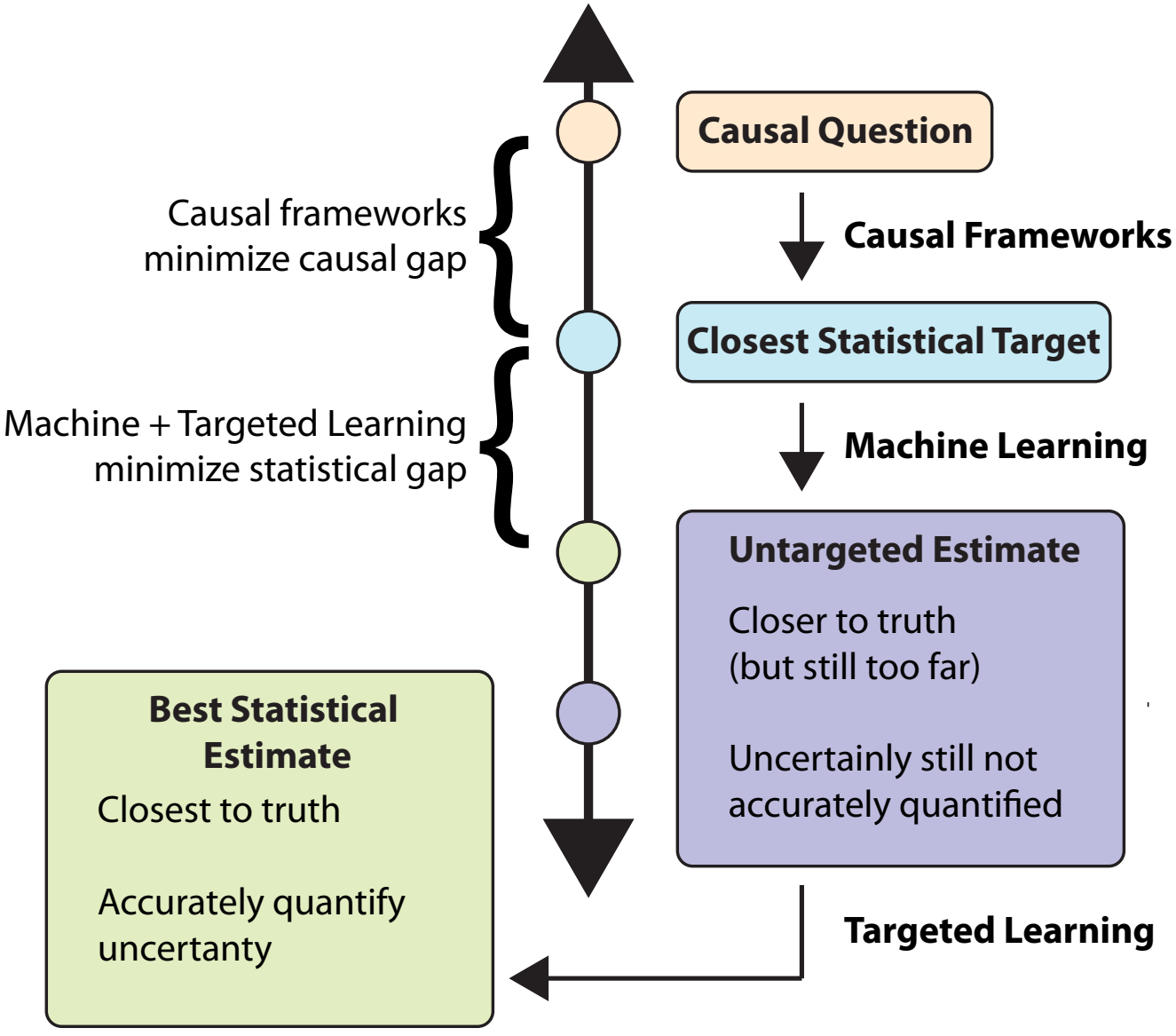
The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

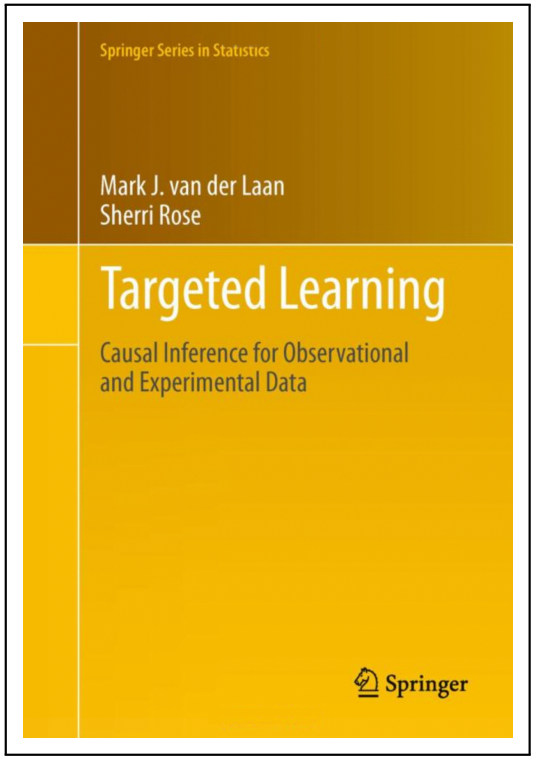
Andrew Gelman and Eric Loken

Outline

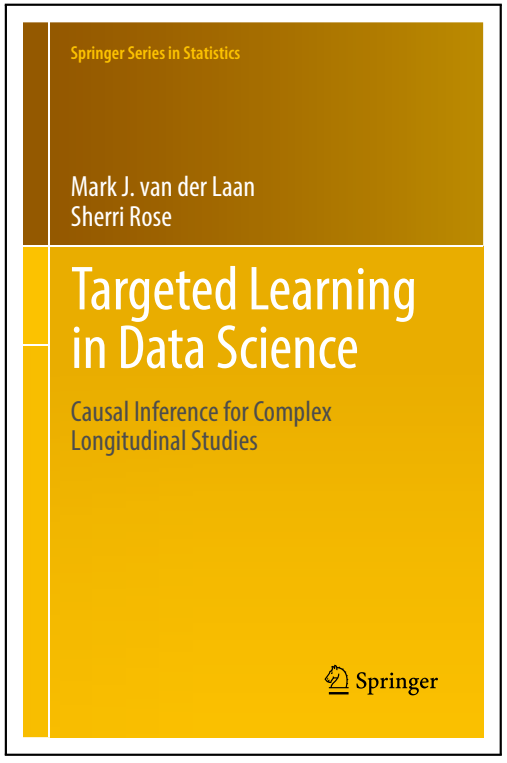
Targeted Learning: Causal Inference Meets Ensemble Machine Learning



Targeted Learning is a subfield of statistics



van der Laan & Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer, 2011.



van der Laan & Rose, *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. New York: Springer, 2018.

`https://vanderlaan-lab.org`

Outline

Roadmap for Targeted Learning

1. Describe observed data
2. Specify statistical model
3. Define statistical query (e.g., using causal roadmap)
4. Construct estimator
5. Obtain inference

Roadmap for Targeted Learning

STEP 1: DESCRIBE OBSERVED DATA

STEP 2: SPECIFY STATISTICAL MODEL

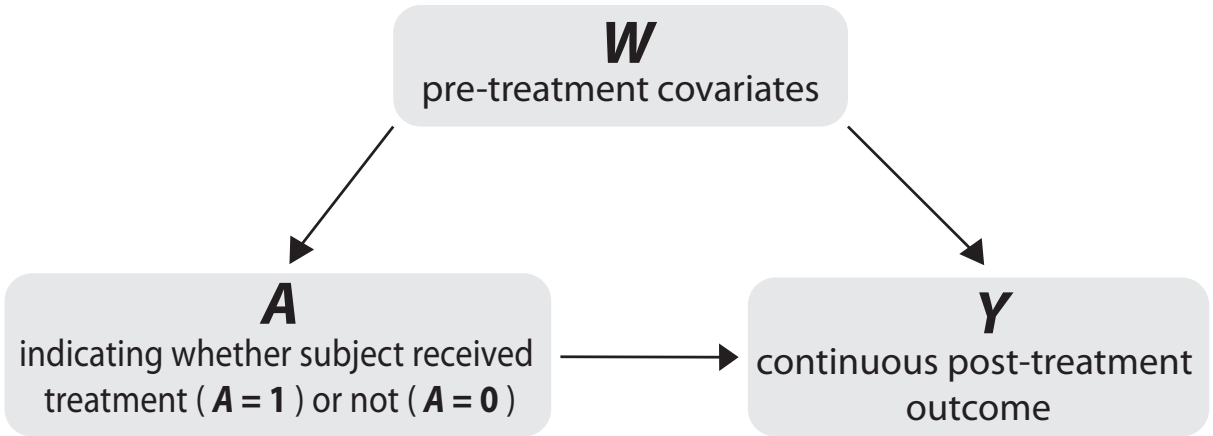
STEP 3: DEFINE STATISTICAL QUERY

STEP 4: CONSTRUCT ESTIMATOR

STEP 5: OBTAIN INFERENCE

$n = 100$ subjects were sampled independently from each other and from the same population distribution P_0

For each subject, pre-treatment covariates (W), treatment (A), and outcome (Y) vectors were measured



Roadmap for Targeted Learning

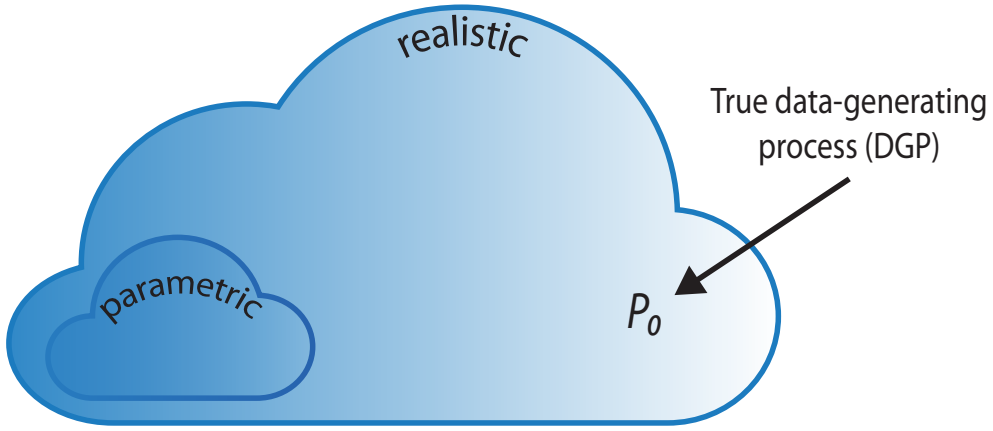
STEP 1:
DESCRIBE
OBSERVED DATA

STEP 2:
SPECIFY
STATISTICAL MODEL

STEP 3:
DEFINE
STATISTICAL QUERY

STEP 4:
CONSTRUCT
ESTIMATOR

STEP 5:
OBTAIN INFERENCE



Standard Approach

Parametric statistical model

Does not contain P_0, the DGP (i.e., misspecified model)

Targeted Learning

Realistic semiparametric or nonparametric statistical model

Defined to ensure P_0 is contained in model

Roadmap for Targeted Learning

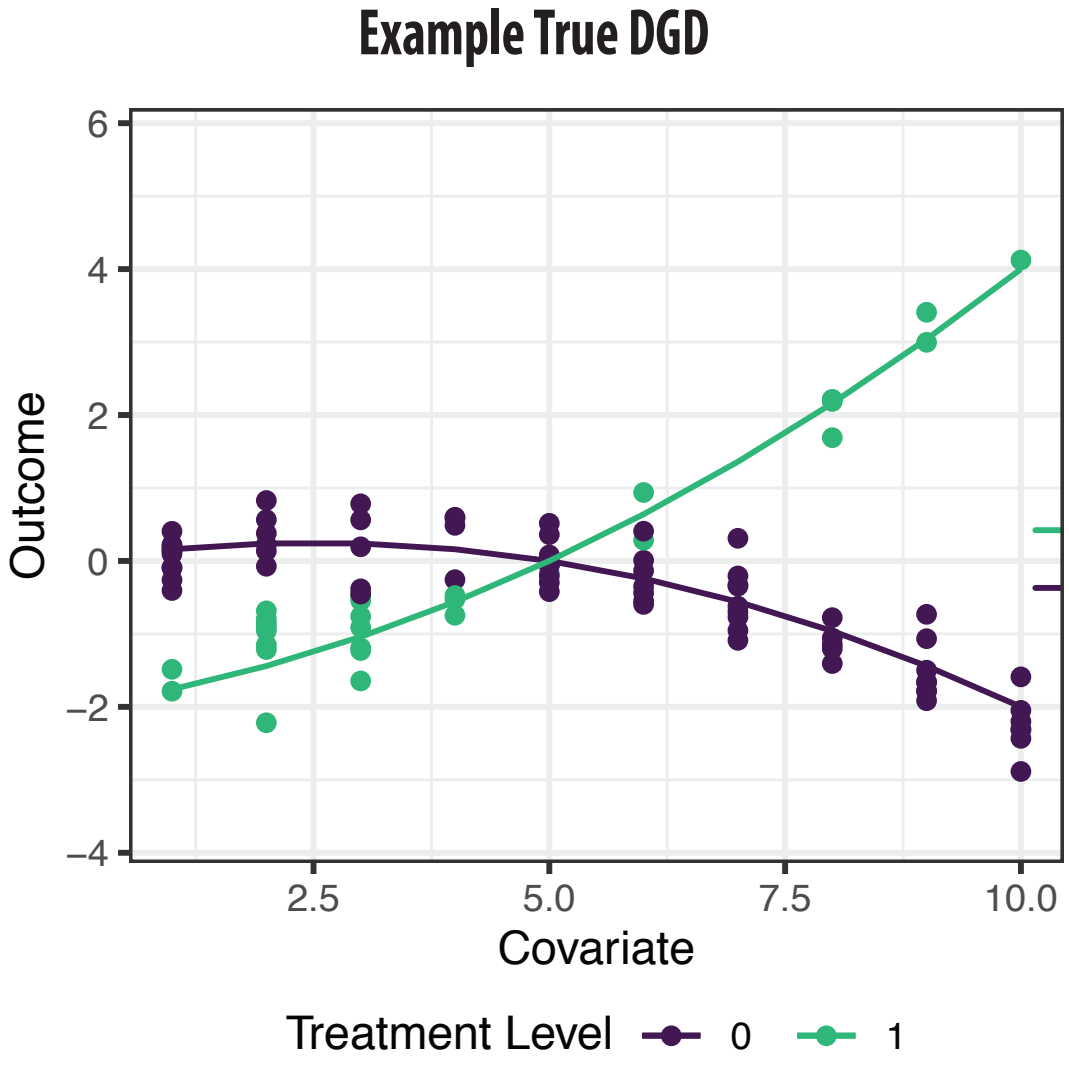
STEP 1:
DESCRIBE
OBSERVED DATA

STEP 2:
SPECIFY
STATISTICAL MODEL

STEP 3:
DEFINE
STATISTICAL QUERY

STEP 4:
CONSTRUCT
ESTIMATOR

STEP 5:
OBTAIN INFERENCE



Roadmap for Targeted Learning

STEP 1:
DESCRIBE
OBSERVED DATA

STEP 2:
SPECIFY
STATISTICAL MODEL

STEP 3:
DEFINE
STATISTICAL QUERY

STEP 4:
CONSTRUCT
ESTIMATOR

STEP 5:
OBTAIN INFERENCE

What is the average difference in outcomes between treatment groups when adjusting for covariates?

$$\Psi(P_0) = E_0\big(E_0[Y|A=1,W] - E_0[Y|A=0,W]\big)$$

Ψ is a function that takes as input P_0
and outputs the answer to the question of interest

The ***assumption of positivity*** is requied to estimate of this quantity from the data. That is, it must be possible to observe both levels of treatment for all strata of *W*.

Additional assumptions are required to interpret this estimand as causal

Causal roadmap for obtaining statistical query answering causal question

Step 3 can be carried out using following causal roadmap:

- Define **potential outcomes** Y_0, Y_1 for each subject, representing (counterfactual) outcome we would have seen if subject would have taken treatment 0 and 1, respectively.
- Link desired full-data (W, Y_0, Y_1) to observed data $O = (W, A, \mathbf{Y} = \mathbf{Y_A})$.
- Define **causal quantity** of interest: $E(Y_1 - Y_0)$, called average treatment effect.
- Establish **identification from DGD**: If treatment is independent of potential outcomes, given W, and positivity holds, then $E_0(Y_1 - Y_0)$ equals target estimand $\Psi(P_0)$.

Roadmap for Targeted Learning

STEP 1:
DESCRIBE
OBSERVED DATA

STEP 2:
SPECIFY
STATISTICAL MODEL

STEP 3:
DEFINE
STATISTICAL QUERY

STEP 4:
CONSTRUCT
ESTIMATOR

STEP 5:
OBTAIN INFERENCE

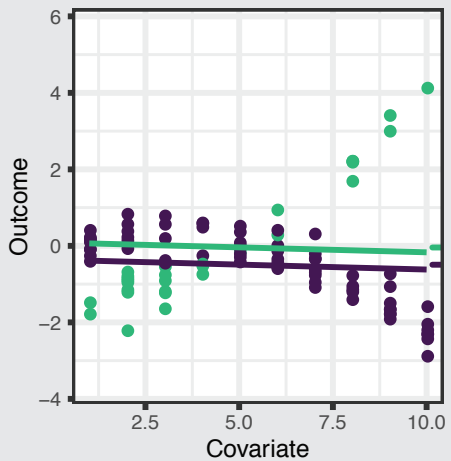
Standard Approach

Generalized Linear Model (GLM)
to estimate

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{A} + \beta_2 \mathbf{W} + \boldsymbol{\epsilon}$$

Estimated coefficients
are biased

Cannot detect heterogeneity
in treatment effect

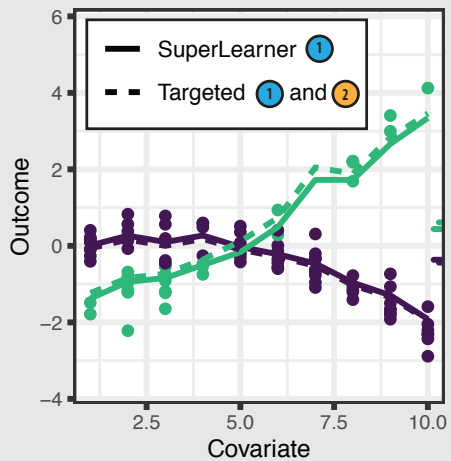


Targeted Learning

TMLE implements
a two-step procedure

1. initial estimation of $E_0[Y|A,W]$ with super (machine) learning
2. targeting towards optimal bias-variance trade-off for $\Psi(P_0)$

TMLE estimates are unbiased
and doubly robust



Roadmap for Targeted Learning

STEP 1:
DESCRIBE
OBSERVED DATA

STEP 2:
SPECIFY
STATISTICAL MODEL

STEP 3:
DEFINE
STATISTICAL QUERY

STEP 4:
CONSTRUCT
ESTIMATOR

STEP 5:
OBTAIN
INFERENCE

Standard Approach

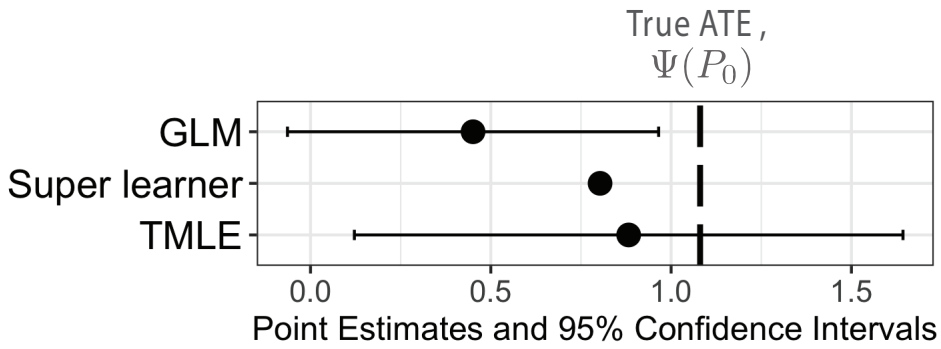
Inference (such as *p*-value and confidence interval) assumes parameteric model is true

Inference is misleading and erroneous

Targeted Learning

Targeting (step 2) improves estimate and makes inference possible

Trustworthy inference obtained with efficient influence function



Outline

Relative Performance Studies of Targeted Learning Estimators

Authors	Title	year	Pro/Con
Rose and van der Laan	Simple Optimal Weighting of Cases and Controls in Case-Control Studies	2008	Pro
Moore and van der Laan	Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation	2009	Pro
Gruber and van der Laan	An application of collaborative targeted maximum likelihood estimation in causal inference and genomics	2010	Pro
Stitelman and van der Laan	Collaborative Targeted Maximum Likelihood for Time to Event Data	2010	Pro
Porter, et al.	The relative performance of targeted maximum likelihood estimators	2011	Pro
Wang, et al.	Finding Quantitative Trait Loci Genes with Collaborative Targeted Maximum Likelihood Learning	2011	Pro
Muñoz and van der Laan	Population Intervention Causal Effects Based on Stochastic Interventions	2011	Neutral
	Diagnosing and responding to violations in the positivity assumption	2012	Pro
van der Laan and Gruber	Targeted Minimum Loss Based Estimation of Causal Effects of Multiple Time Point Interventions	2012	Pro
Gruber and van der Laan	An Application of Targeted Maximum Likelihood Estimation to the Meta-Analysis of Safety Data	2013	Neutral
Lendle, et al.	Targeted maximum likelihood estimation in safety analysis	2013	Pro
Díaz and van der Laan	Targeted Data Adaptive Estimation of the Causal Dose Response Curve	2013	Pro
Schnitzer, et al.	time-dependent treatment effects under density misspecification	2013	Neutral
Kreif, et al.	Misspecification: a comparison of targeted maximum likelihood estimation with bias-corrected matching.	2014	Pro
Schnitzer, et al.	Effect of breastfeeding on gastrointestinal infection in infants: A targeted maximum likelihood approach for clustered longitudinal data	2014	Pro
Pang, et al.	Effect Estimation in Point-Exposure Studies with Binary Outcomes and High-Dimensional Covariate Data--A Comparison of Targeted Maximum Likelihood Estimation and Inverse Probability of Treatment Weighting	2016	Pro
Pang, et al.	Effect Estimation in Point-Exposure Studies with Binary Outcomes and High-Dimensional Covariate Data--A Comparison of Targeted Maximum Likelihood Estimation and Inverse Probability of Treatment Weighting	2016	Pro
Schnitzer, et al.	Variable Selection for Confounder Control, Flexible Modeling and Collaborative Targeted Minimum Loss-Based Estimation in Causal Inference	2016	Pro
Zheng, et al.	Doubly Robust and Efficient Estimation of Marginal Structural Models for the Hazard Function	2016	Pro
Schnitzer, et al.	Double robust and efficient estimation of a prognostic model for events in the presence of dependent censoring	2016	Pro
Schuler and Rose	Targeted maximum likelihood estimation for causal inference in observational studies	2017	Pro
Gruber and van der Laan	Collaborative Targeted Maximum Likelihood Estimation to Assess Causal	2018	Pro
Pirracchio, et al.	Collaborative targeted maximum likelihood estimation for variable importance measure: Illustration for functional outcome prediction in mild traumatic brain injuries	2018	Pro
Luque-Fernandez, et al.	Targeted maximum likelihood estimation for a binary treatment: A tutorial	2018	Pro
Levy, et al.	A Fundamental Measure of Treatment Effect Heterogeneity	2018	Pro
Ju, et al.	dimensional data	2019	Pro
Ju, et al.	On adaptive propensity score truncation in causal inference	2019	Pro
Bahamyirou, et al.	using machine learning methods with doubly robust causal estimators	2019	Pro
Wei, et al.	A Data-Adaptive Targeted Learning Approach of Evaluating Viscoelastic Assay Driven Trauma Treatment Protocols	2019	Pro
Rudolph, et al.	Complier Stochastic Direct Effects: Identification and Robust Estimation	2019	Pro
Chatton, et al.	targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study	2020	Con
Talbot an Beaudoin	A generalized double robust Bayesian model averaging approach to causal effect estimation with application to the Study of Osteoporotic Fractures	2020	Con

Improving upon the current standard of predictive analytics in the ICU

THE LANCET
Respiratory Medicine
Volume 3, Issue 1, January 2015, Pages 42-52

Articles

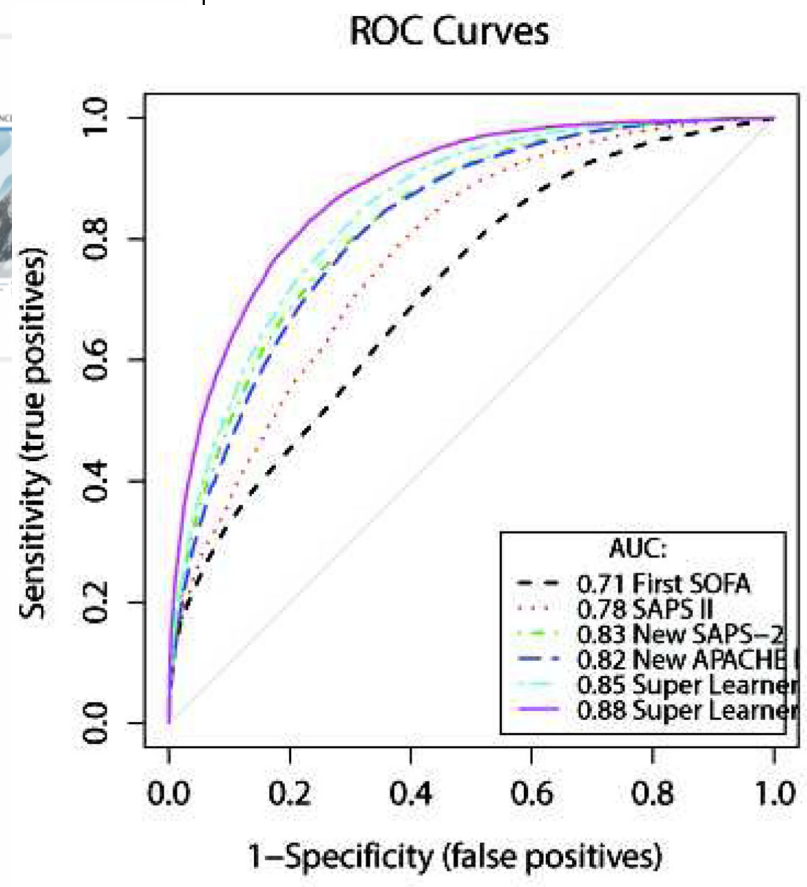
Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study

Improving upon the current standard of predictive analytics in the ICU

THE LANCET
Respiratory Medicine
Volume 3, Issue 1, January 2015, Pages 42-52

Articles

Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study



Estimating the causal effect of a community-level intervention in a clustered RCT

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

HIV Testing and Treatment with the Use of a Community Health Approach in Rural Africa

D.V. Havlir, L.B. Balzer, E.D. Charlebois, T.D. Clark, D. Kwarisiima, J. Ayieko, J. Kabami, N. Sang, T. Liegler, G. Chamie, C.S. Camlin, V. Jain, K. Kadede, M. Atukunda, T. Ruel, S.B. Shade, E. Ssemmondo, D.M. Byonanebye, F. Mwangwa, A. Owaraganise, W. Ololo, D. Black, K. Snyman, R. Burger, M. Getahun, J. Achando, B. Awuonda, H. Nakato, J. Kironde, S. Okiror, H. Thirumurthy, C. Koss, L. Brown, C. Marquez, J. Schwab, G. Lavoy, A. Plenty, E. Mugoma Wafula, P. Omanya, Y.-H. Chen, J.F. Rooney, M. Bacon, M. van der Laan, C.R. Cohen, E. Bukusi, M.R. Kamya, and M. Petersen

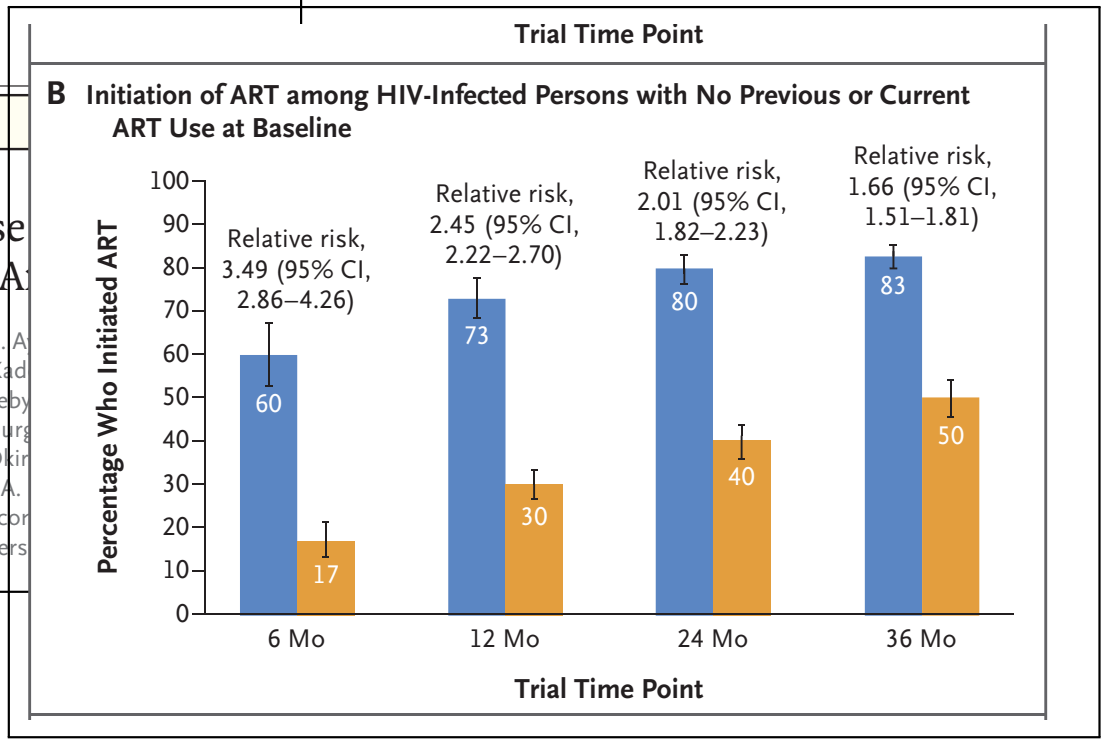
Estimating the causal effect of a community-level intervention in a clustered RCT

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

HIV Testing and Treatment with the Use
Community Health Approach in Rural A

D.V. Havlir, L.B. Balzer, E.D. Charlebois, T.D. Clark, D. Kwarisiima, J. A
J. Kabami, N. Sang, T. Liegler, G. Chamie, C.S. Camlin, V. Jain, K. Kad
M. Atukunda, T. Ruel, S.B. Shade, E. Ssemmondo, D.M. Byonaneby
F. Mwangwa, A. Owaraganise, W. Olilo, D. Black, K. Snyman, R. Burg
M. Getahun, J. Achando, B. Awuonda, H. Nakato, J. Kironde, S. Okir
H. Thirumurthy, C. Koss, L. Brown, C. Marquez, J. Schwab, G. Lavoy, A.
E. Mugoma Wafula, P. Omanya, Y.-H. Chen, J.F. Rooney, M. Bacor
M. van der Laan, C.R. Cohen, E. Bukusi, M.R. Kamya, and M. Peters



Estimating the impact of genetic polymorphisms on the efficacy of malaria vaccine on the time to infection

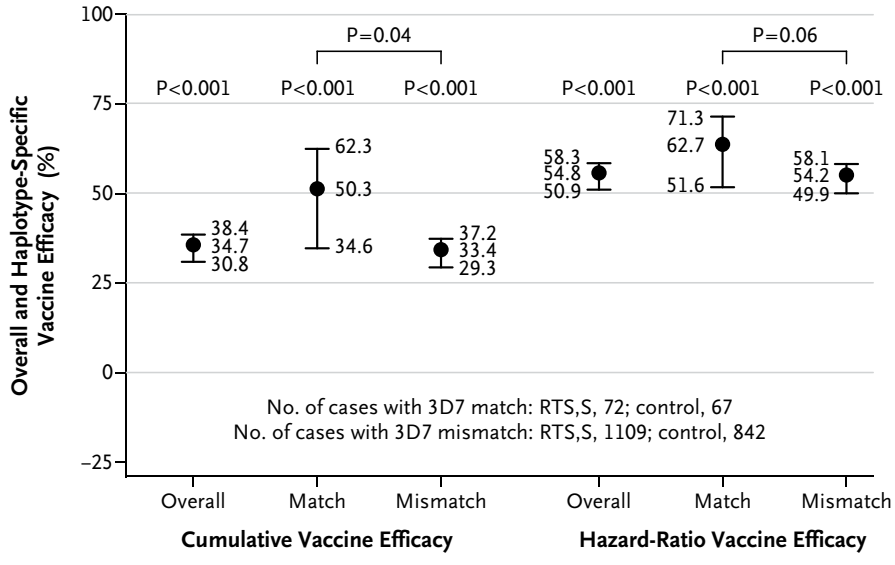
The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine

D.E. Neafsey, M. Juraska, T. Bedford, D. Benkeser, C. Valim, A. Griggs, M. Lievens, S. Abdulla, S. Adjei, T. Agbenyega, S.T. Agnandji, P. Aide, S. Anderson, D. Ansong, J.J. Aponte, K.P. Asante, P. Bejon, A.J. Birkett, M. Bruls, K.M. Connolly, U. D'Alessandro, C. Dobaño, S. Gesase, B. Greenwood, J. Grimsby, H. Tinto, M.J. Hamel, I. Hoffman, P. Kamthunzi, S. Kariuki, P.G. Kremsner, A. Leach, B. Lell, N.J. Lennon, J. Lusingu, K. Marsh, F. Martinson, J.T. Molel, E.L. Moss, P. Njuguna, C.F. Ockenhouse, B. Ragama Ogutu, W. Otieno, L. Otieno, K. Otieno, S. Owusu-Agyei, D.J. Park, K. Pellé, D. Robbins, C. Russ, E.M. Ryan, J. Sacarlal, B. Sogoloff, H. Sorgho, M. Tanner, T. Theander, I. Valea, S.K. Volkman, Q. Yu, D. Lapierre, B.W. Birren, P.B. Gilbert, and D.F. Wirth

D Cumulative and Hazard-Ratio Vaccine Efficacy



Estimating the cumulative, long-term impacts of environmental exposures

ORIGINAL ARTICLE

OPEN

Occupational Exposure to $PM_{2.5}$ and Incidence of Ischemic Heart Disease

Longitudinal Targeted Minimum Loss-based Estimation

Daniel M. Brown,[a] *Maya Petersen,*[a] *Sadie Costello,*[b] *Elizabeth M. Noth,*[b] *Katherine Hammond,*[b] *Mark Cullen,*[c] *Mark van der Laan,*[a] *and Ellen Eisen*[b]

Estimating the cumulative, long-term impacts of environmental exposures

ORIGINAL ARTICLE

OPEN

Occupational Exposure to $PM_{2.5}$ and Ischemic Heart Disease

Longitudinal Targeted Minimum Loss-bas

Daniel M. Brown,[a] Maya Petersen,[a] Sadie Costello,[b] Elizabeth M. Noti
Mark Cullen,[c] Mark van der Laan,[a] and Ellen Ei

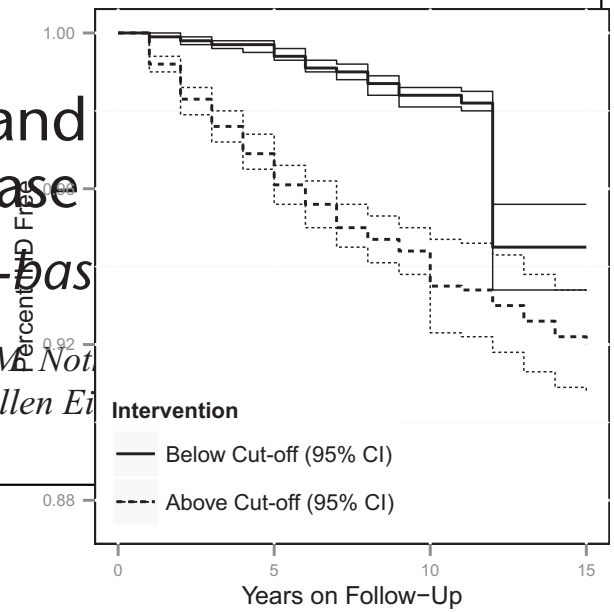


FIGURE 4. Estimated cumulative survival from IHD and 95% confidence intervals, adjusted for measured baseline and time-varying risk factors, among the fabricator worker population if continuously exposed versus unexposed at the 10th percentile cut-off of 0.06 mg/m^3. The *thick lines* represent the point estimates, whereas the *thinner lines* represent the 95% confidence intervals, with the line types (*dotted* vs. *solid*) indicating if the estimate is for above or below the cut-off.

Comparing strategies for diabetes treatment intensification in Comparative Effectiveness Research (CER) study

Statistics
in Medicine

Research Article

Received 24 May 2013, Accepted 5 January 2014 Published online 17 February 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6099

Targeted learning in real-world comparative effectiveness research with time-varying interventions

Romain Neugebauer,[a*†] Julie A. Schmittdiel[a] and Mark J. van der Laan[b]

Comparing strategies for diabetes treatment intensification in Comparative Effectiveness Research (CER) study

Statistics
in Medicine

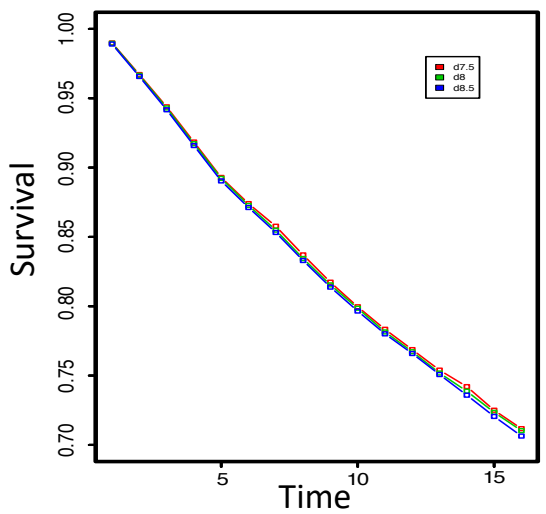
Research Article

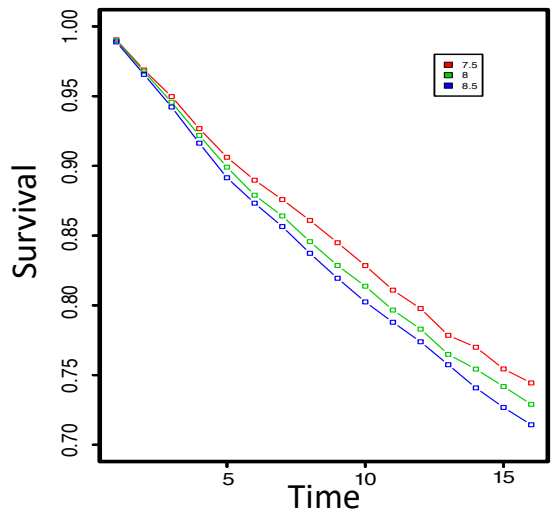
Received 24 May 2013, Accepted 5 January 2014 Published online 17 February 2014 in Wiley Online Library

(wileyonlinelibrary.com)

Targeted comparat time-vary

Romain Neugel Mark J. van de





Standard methods: No benefit to more aggressive intensification strategy

Targeted Learning: More aggressive intensification protocols result in better outcomes

Identifying contributing factors for health care spending

DOI: 10.1111/1475-6773.12848
METHODS ARTICLE

Robust Machine Learning Variable Importance Analyses of Medical Conditions for Health Care Spending

Sherri Rose

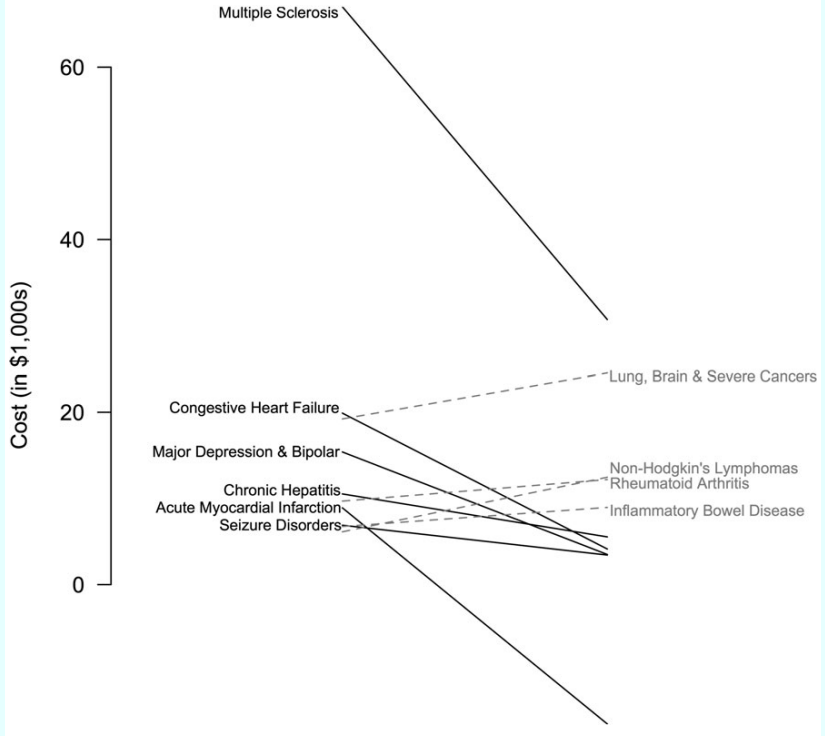
Identifying contributing factors for health care spending

DOI: 10.1111/1475-6773.12848
METHODS ARTICLE

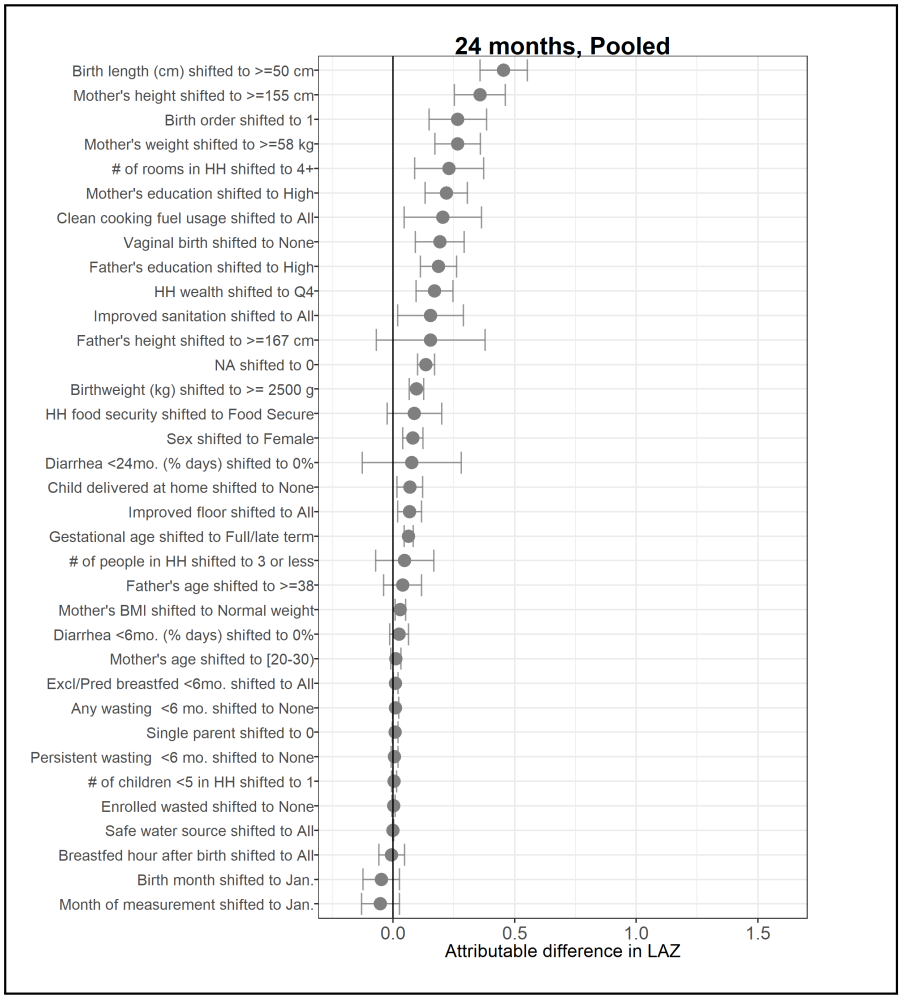
Robust Machine Learning Vari
Importance Analyses of Medic
Conditions for Health Care Spe

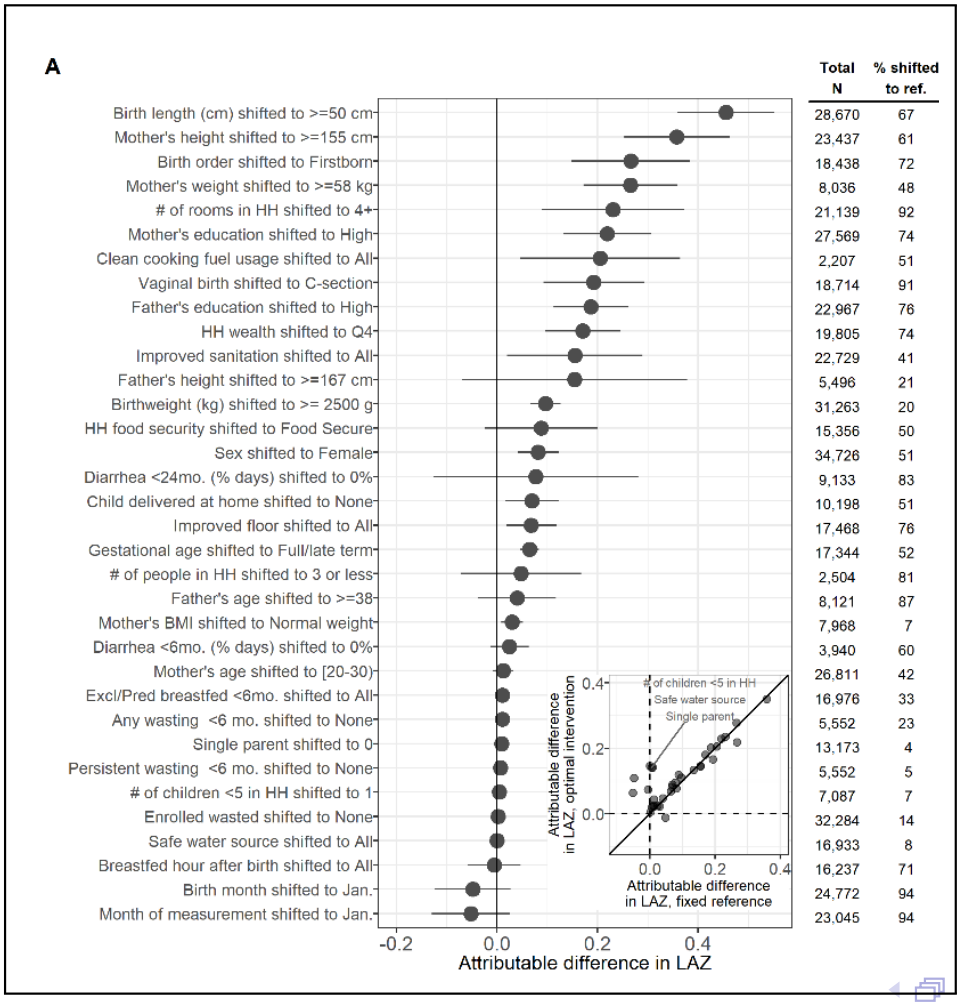
Sherri Rose

Figure 4: Top 10 Largest Targeted Learning Effect Estimates



Identifying contributing factors of childhood mortality and estimating their impact in WASH benefits observational study





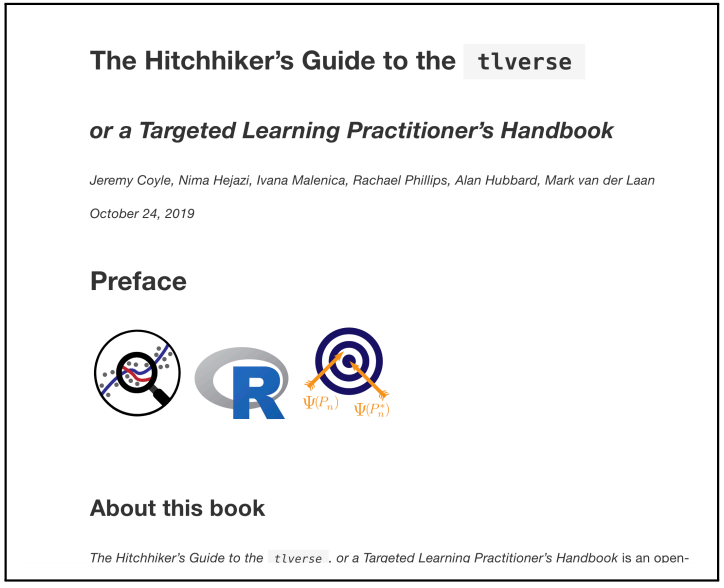
Outline

`tlverse` - Targeted Learning software ecosystem in `R`

- A curated collection of `R` packages for Targeted Learning
- Shares a consistent underlying philosophy, grammar, and set of data structures
- Open source
- Designed for generality, usability, and extensibility

`tlverse` outreach to train and support practitioners

- May 2019 - Atlantic Causal Inference Conference (ACIC) Workshop
- June 2019 - `tlverse` book →
- October 2019 - University of Pittsburgh School of Public Health Workshop
- November 2019 - Bill & Melinda Gates Foundation Workshop
- December 2019 - Deming Conference on Applied Statistics Workshop



The Hitchhiker's Guide to the `tlverse`

or a Targeted Learning Practitioner's Handbook

Jeremy Coyle, Nima Hejazi, Ivana Malenica, Rachael Phillips, Alan Hubbard, Mark van der Laan

October 24, 2019

Preface

About this book

The Hitchhiker's Guide to the `tlverse`. or a Targeted Learning Practitioner's Handbook is an open-

- February 2020 - Conference on Statistical Practice (CSP) Workshop
- March 2020 - Alan Turing Institute Workshop

Outline

Partial list of available functions/applications in Targeted Learning

- R framework for targeted learning (`https://github.com/tlverse`)
- Longitudinal intervention effects (`https://cran.r-project.org/web/packages/ltmle/index.html`)
- Estimate impact of specific pathways (mediation effects; `https://github.com/tlverse/tmle3mediate`).
- Stochastic interventions (e.g., shifting exposure by some amount; `https://github.com/tlverse/tmle3shift`)
- Predicting differential treatment effects. Using this to derive optimal treatment rules for precision health (`https://github.com/tlverse/tmle3mopttx`)
- Towards truly optimal machine learning via highly adaptive lasso or HAL (`https://github.com/tlverse/hal9001`).
- Data adaptive parameters (using the data to discover the interesting parameter and still derive inference (`https://github.com/ck37/varImpact`)
- Constrained machine learning (e.g., for calibration).
- Effects of mixtures (e.g., mixtures of toxicants on health).
- Super Learning for optimally weighted average of multivariate outcomes (`https://github.com/benkeser/r2weight`)
- Online Superlearning for streaming data (e.g., waveforms, time-series of Covid 19 data,) Adaptive surveillance/designs (learning from the past to more efficiently assignment treatment or recruit individuals). Etc.

Concluding Remarks

- **Targeted Learning** *optimally estimates* the causal impact of an intervention on an outcome for complex real-world data.
- It integrates **causal inference**, **machine learning**, **statistical theory**.
- Targeted Learning learns better answers to causal, actionable questions which result in improved policy, treatments, etc.
- The estimate is accompanied with accurate quantification of uncertainty such as **confidence interval and p-value**.
- We have developed an ongoing targeted learning software environment `tlverse` with growing number of tools and tutorials.