The `tlverse` Software Ecosystem for Targeted Learning

Short-Course at the Conference of Statistical Practice

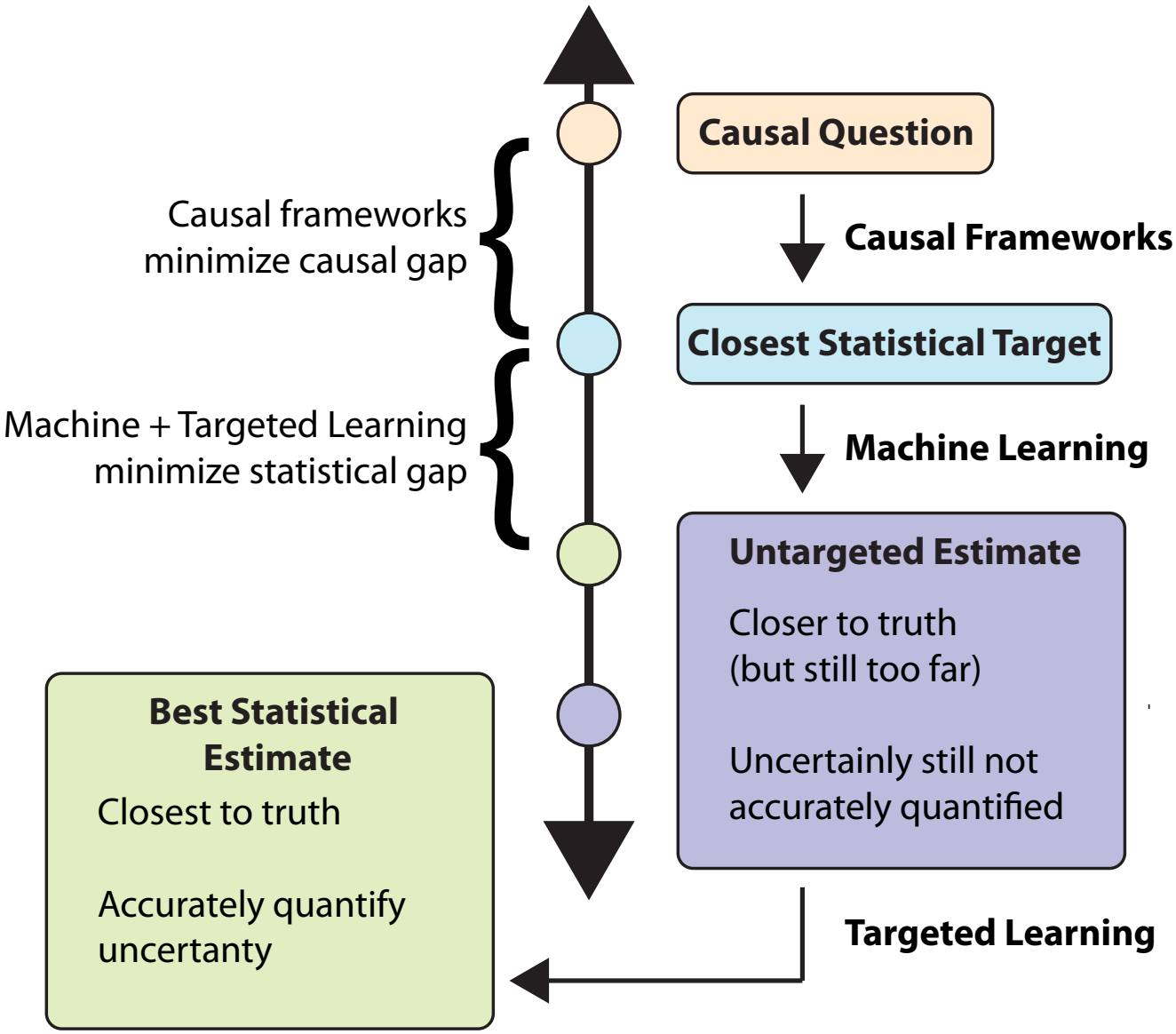
Mark van der Laan Alan Hubbard Jeremy Coyle Nima Hejazi
Ivana Malenica Rachael Phillips

Group in Biostatistics
University of California at Berkeley

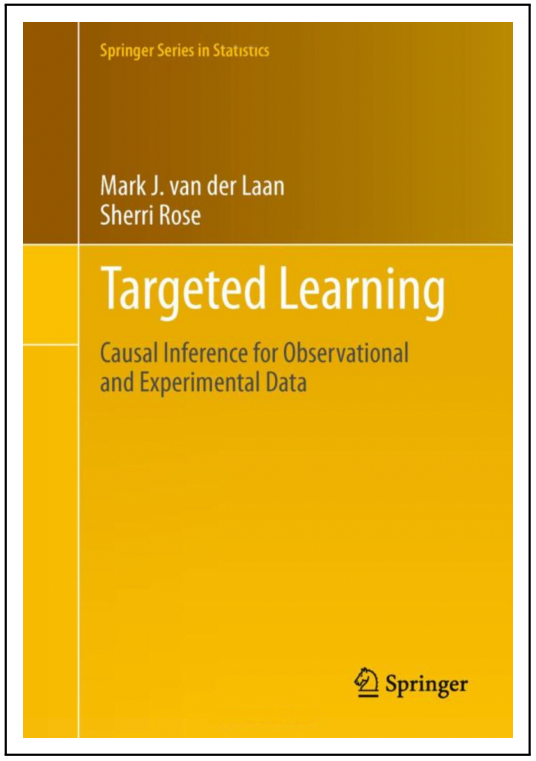
February 20, 2020
8:00A-5:30P

Outline

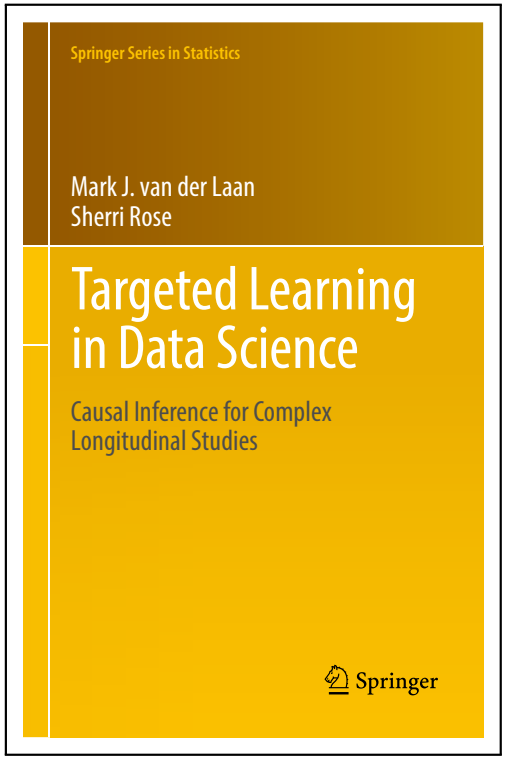
Targeted Learning fills a much needed gap in machine learning and causal inference



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

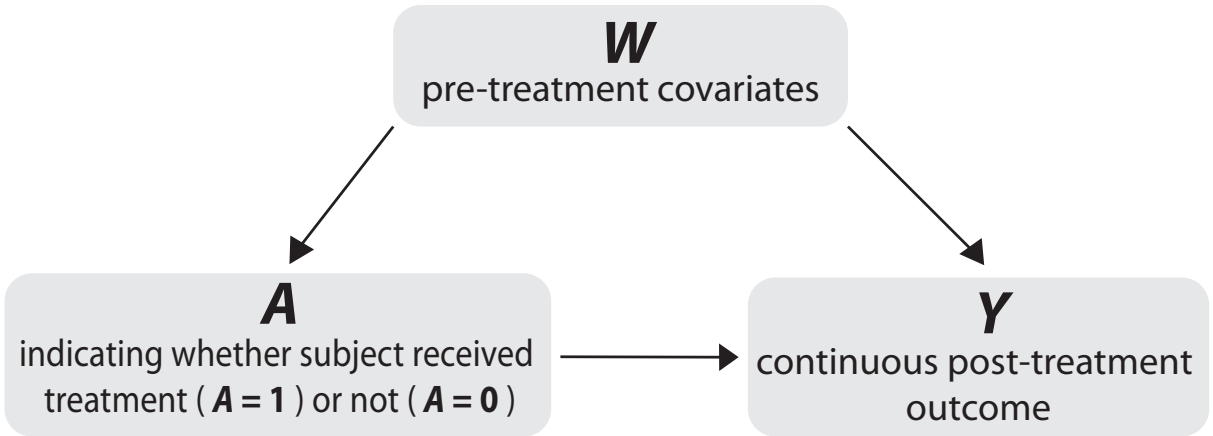
STEP 3: DEFINE STATISTICAL QUERY

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

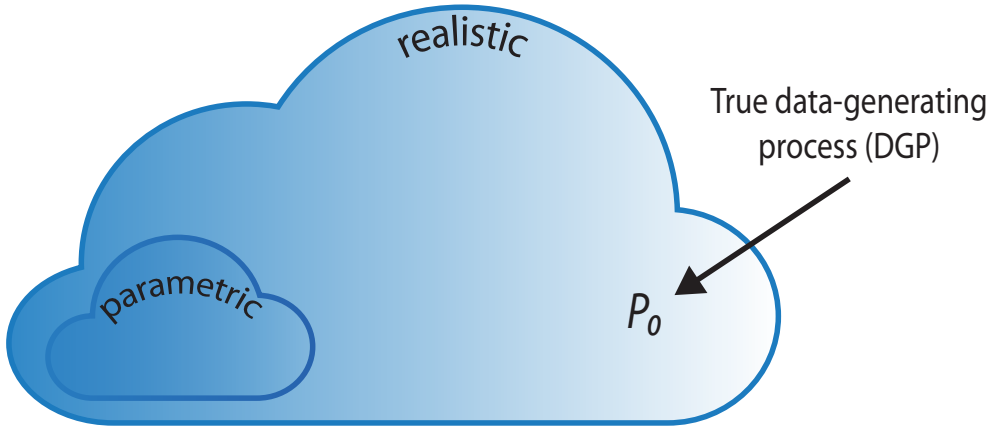
STEP 1:
DESCRIBE
OBSERVED DATA

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

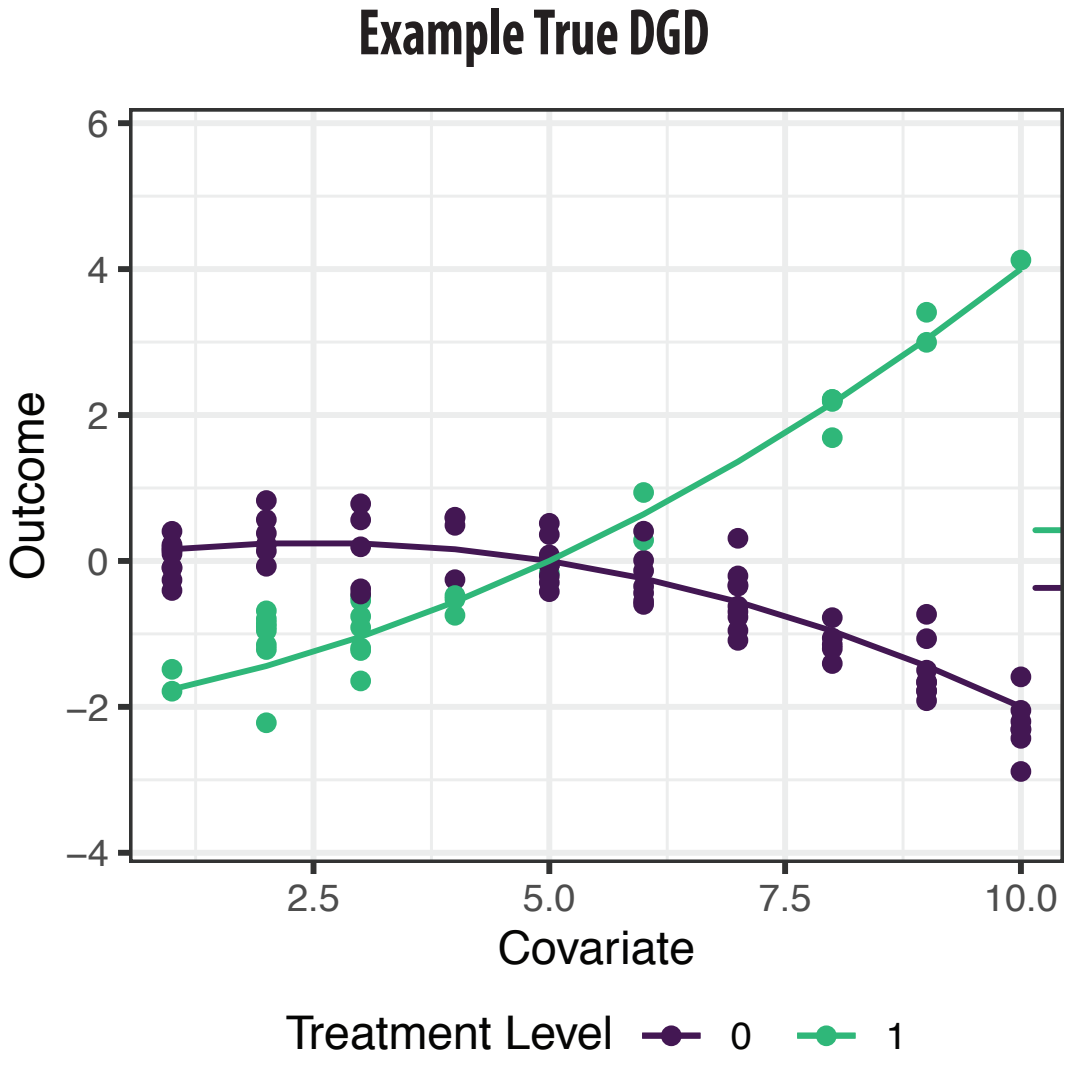
STEP 1:
DESCRIBE
OBSERVED DATA

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 DGP**: 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

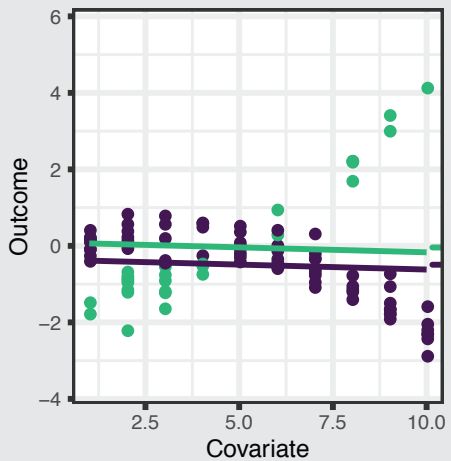
Standard Approach

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

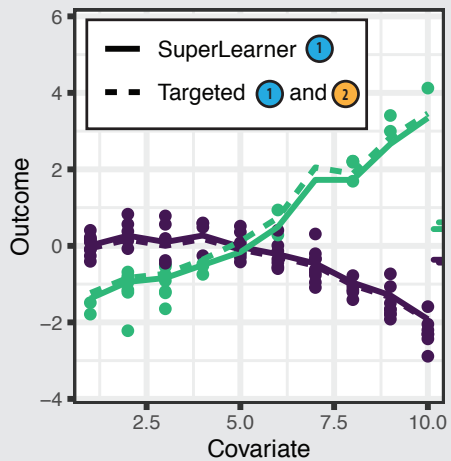


Targeted Learning

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

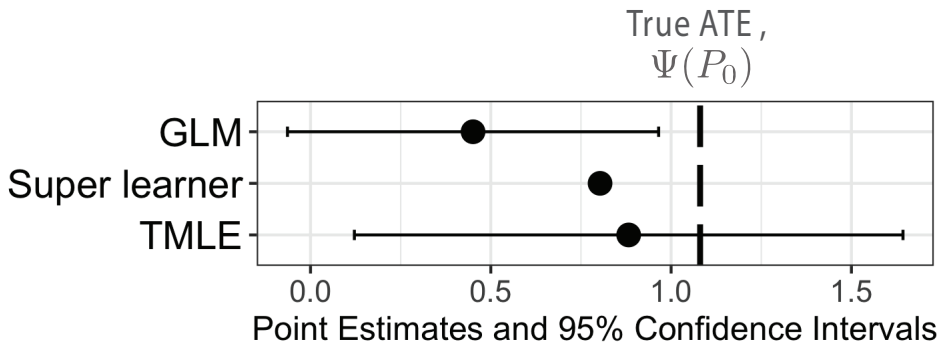
Inference (such as *p*-value and confidence interval) assumes parameteric model is true

Inference is misleading and erroneous

Targeted Learning

Targeting (step 2) improves estimate and makes inference possible

Trustworthy inference obtained with efficient influence function



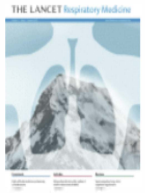
Roadmap of Statistical Learning Summary

1. **Describe observed data**: Realization of a random variable $O^n = (O_1, \ldots, O_n)$ with a probability distribution (say) P_0^n, indexed by "sample size" n.
2. **Specify statistical model of stochastic system of observed data realistically**: Statistical model $\mathcal{M}^n$ is set of possible probability distributions of the data.
3. **Define query about stochastic system**: Function Ψ from model $\mathcal{M}^n$ to real line, where $\Psi(P_0^n)$ is the true answer to query about our stochastic system.
4. **Construct estimator**: An a priori-specified algorithm that takes the observed data O^n and returns an estimate ψ_n to the *true answer to query*. Benchmarked by a dissimilarity-measure (e.g., MSE) w.r.t true answer to query.
5. **Obtain inference**: Establish approximate sampling probability distribution of the estimator (e.g., based on CLT), and corresponding statistical inference, such as the confidence interval for the true answer to query.

Outline

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

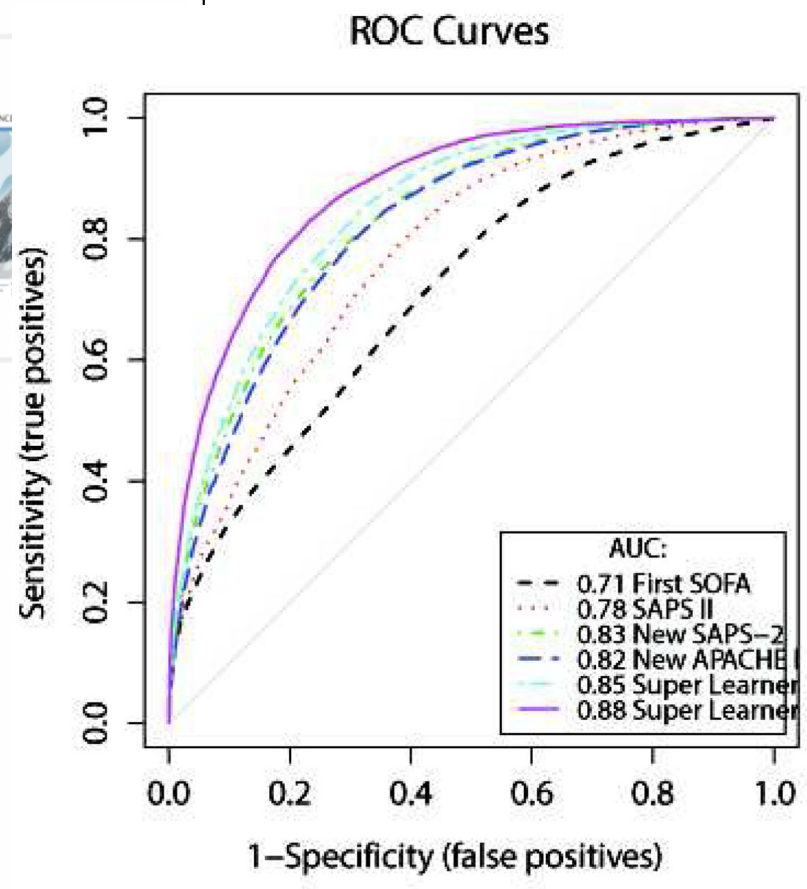
Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study

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



Average treatment effect in an observational study

International Archives of Occupational and Environmental Health (2019) 92:629–638
https://doi.org/10.1007/s00420-018-1397-1

ORIGINAL ARTICLE

An educational intervention to improve knowledge about prevention against occupational asthma and allergies using targeted maximum likelihood estimation

Daloha Rodríguez-Molina[1,2] · Swaantje Barth[1] · Ronald Herrera[1] · Constanze Rossmann[3] · Katja Radon[1] · Veronika Karnowski[4]

Received: 15 March 2018 / Accepted: 13 December 2018 / Published online: 14 January 2019

Average treatment effect in an observational study

International Archives of Occupational and Environmental Health (2019) 92:629–638
https://doi.org/10.1007/s00420-018-1397-1

ORIGINAL ARTICLE

An educational intervention to improve knowledge abo against occupational asthma and allergies using targete likelihood estimation

Daloha Rodríguez-Molina[1,2] · Swaantje Barth[1] · Ronald Herrera[1] · Constanze R Veronika Karnowski[4]

Received: 15 March 2018 / Accepted: 13 December 2018 / Published online: 14 January 2019

Table 4 Adjusted average treatment effects of the intervention ($n=116$), Bavaria, Germany, 2014

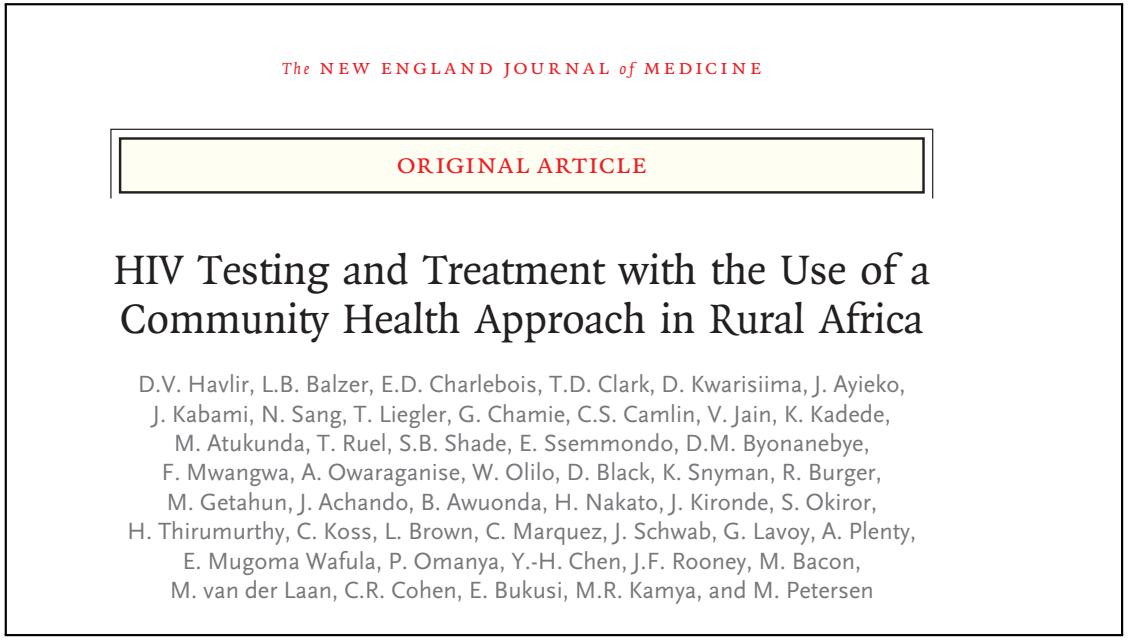
	All six correct measures	At least five correct measures	At least four correct measures
Additive ATE			
Parameter	18.44%	55.53%	29.60%
95% CI	(7.3–29.58%)	(36.96–74.09%)	(12.2–47.0%)
Additive ATT			
Parameter	16.9%	63.07%	62.78%
95% CI	(5.38–28.51%)	(46.02–80.13%)	(41.64–83.93%)
Additive ATC			
Parameter	16.8%	32.28%	18.97%
95% CI	(5.02–28.57%)	(12.84–51.72%)	(1.91–36.02%)

Adjusted for sex, age, education level, smoking status, presence of asthma or rhinoconjunctivitis, riskperception, parental asthma, and knowledge about preventive measures against asthma and allergies before the intervention

The adjusted model using TMLE allowed including both observed data ($n=47$) and missing values ($n=69$) as parameters

ATE average treatment effect, *ATT* average treatment effect on the treated, *CI* confidence interval, *ATC* average treatment effect on the controls, *TMLE* targeted maximum likelihood estimation

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. Olilo, 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

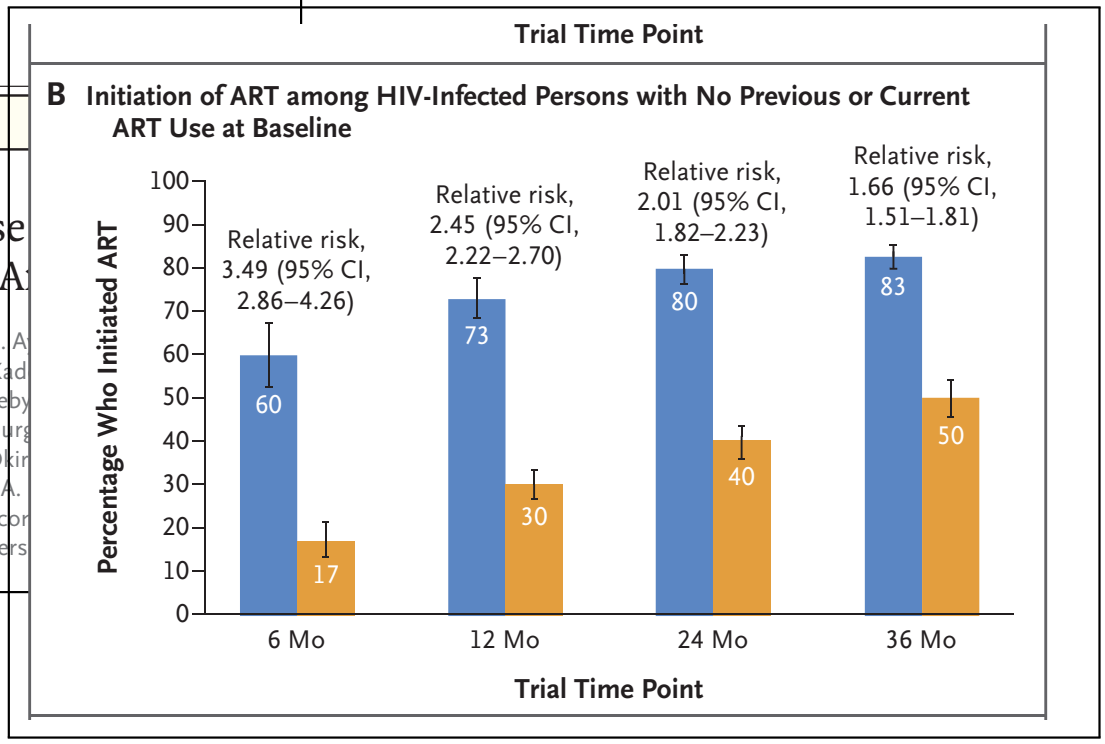
Estimating the causal effect of a community-level intervention in a clustered RCT

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



Increasing precision and accuracy by accounting for missing data in estimating impacts of HIV treatment program in clustered RCT

Research

JAMA | Original Investigation

Association of Implementation of a Universal Testing and Treatment Intervention With HIV Diagnosis, Receipt of Antiretroviral Therapy, and Viral Suppression in East Africa

Maya Petersen, MD, PhD; Laura Balzer, PhD; Dalsone Kwarsiima, MBChB, MPH; Norton Sang, MA; Gabriel Chamie, MD, MPH; James Ayieko, MBChB, MPH; Jane Kabami, MPH; Asiphas Owaraganise, MBChB; Teri Liegler, PhD; Florence Mwangwa, MBChB; Kevin Kadede, MA; Vivek Jain, MD, MAS; Albert Plenty, MS; Lillian Brown, MD, PhD; Geoff Lavoy; Joshua Schwab, MS; Douglas Black, BA; Mark van der Laan, PhD; Elizabeth A. Bukusi, MBChB, PhD; Craig R. Cohen, MD, MPH; Tamara D. Clark, MHS; Edwin Charlebois, MPH, PhD; Moses Kamya, MMed; Diane Havlir, MD

Increasing precision and accuracy by accounting for missing data in estimating impacts of HIV treatment program in clustered RCT

Research

JAMA | Original Investigation

Association of Impleme and Treatment Interve of Antiretroviral Therap

Maya Petersen, MD, PhD; Laura Balzer, PhD; Dalsone Kwarsiima, MBChB, MPH; Norton Sang, MA; Gabriel Chamie, MD, MPH; James Ayieko, MBChB, MPH; Jane Kabami, MPH; Asiphas Owaraganise, MBChB; Teri Liegler, PhD; Florence Mwangwa, MBChB; Kevin Kadede, MA; Vivek Jain, MD, MAS; Albert Plenty, MS; Lillian Brown, MD, PhD; Geoff Lavoy; Joshua Schwab, MS; Douglas Black, BA; Mark van der Laan, PhD; Elizabeth A. Bukusi, MBChB, PhD; Craig R. Cohen, MD, MPH; Tamara D. Clark, MHS; Edwin Charlebois, MPH, PhD; Moses Kamya, MMed; Diane Havlir, MD

Table 2. Postbaseline HIV Viral Suppression in a Closed Cohort of HIV-Positive Stable Residents of 16 SEARCH Intervention Communities in Rural Uganda and Kenya Who Were Diagnosed At or Before Baseline (n = 7108)[a]

Baseline Diagnosis, Treatment, and Suppression Status	No. of HIV-Positive Residents (%)[a]	Follow-up Year 1		Follow-up Year 2	
		No. of Residents With Viral Suppression/Total No. of Residents With Measured HIV RNA (%)[b]	Adjusted Proportion, % (95% CI)[c]	No. of Residents With Viral Suppression/Total No. of Residents With Measured HIV RNA (%)[d]	Adjusted Proportion, % (95% CI)[c]
Overall	7108 (100)	4682/5578 (83.9)	79.7 (78.7-80.8)	4602/5215 (88.2)	83.8 (82.8-84.9)
Newly diagnosed (HIV RNA≥500 copies/mL)	2080 (29.3)	963/1321 (72.9)	62.8 (60.4-65.2)	965/1205 (80.1)	68.8 (66.4-71.2)
Previously diagnosed with no ART (HIV RNA≥500 copies/mL)	990 (13.9)	649/812 (79.9)	78.1 (75.3-80.8)	685/778 (88.0)	86.5 (84.2-88.8)
Previous or current ART	4038 (56.8)	3070/3445 (89.1)	88.8 (87.7-89.9)	2952/3232 (91.3)	90.5 (89.4-91.6)
HIV RNA not measured	1063 (15.0)	732/846 (86.5)	86.6 (84.3-88.9)	685/779 (87.9)	87.2 (84.9-89.5)
HIV RNA≥500 copies/mL	426 (6.0)	175/355 (49.3)	49.5 (44.2-54.7)	204/325 (62.8)	62.2 (57.2-67.2)
HIV RNA<500 copies/mL	2549 (35.9)	2163/2244 (96.4)	96.3 (95.6-97.1)	2063/2128 (96.9)	96.8 (96.0-97.6)

Abbreviations: ART, antiretroviral treatment; HIV, human immunodeficiency ... censoring by death and migration out of the community and for missing HIV

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

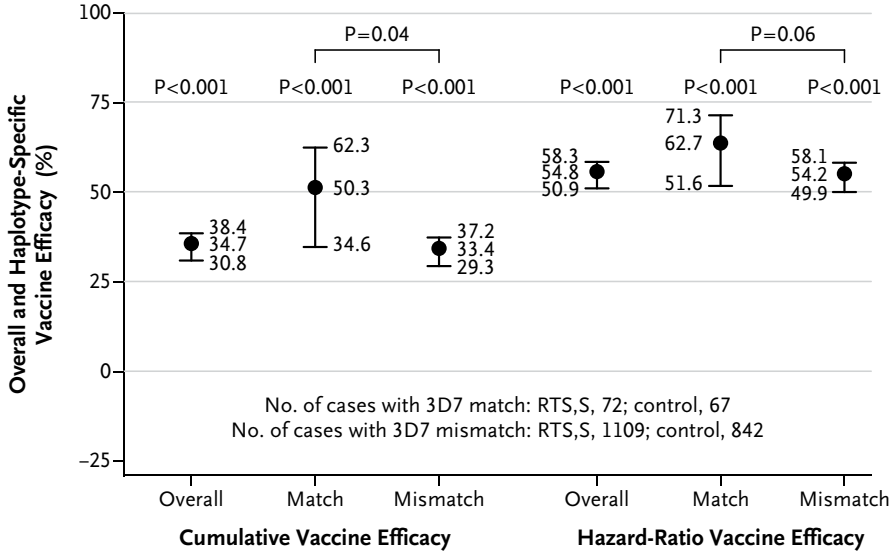
The NEW ENGLAND JOURNAL of MEDICINE

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*

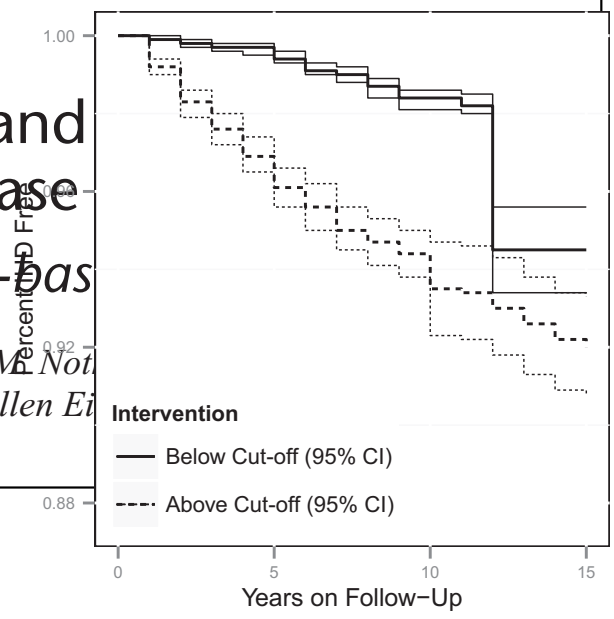


FIGURE 4. Estimated cumulative survival from IHD and 95% confidence intervals, adjusted for measured baseline and time-varying risk factors, among the fabricator worker population if continuously exposed versus unexposed at the 10th percentile cut-off of 0.06 mg/m^3. The *thick lines* represent the point 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

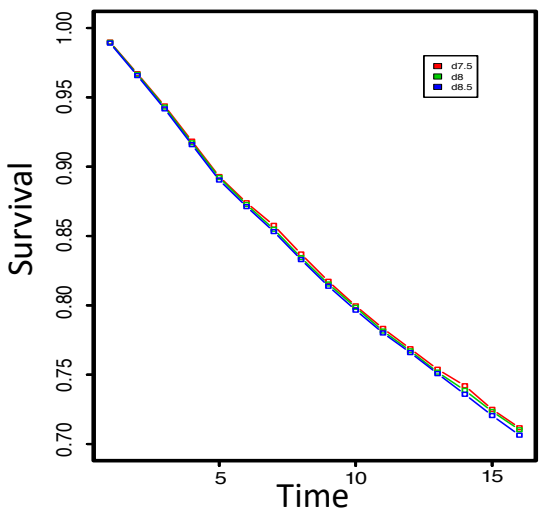
Research Article

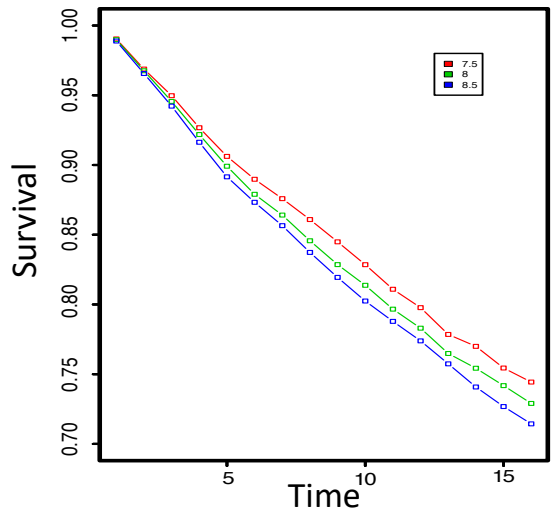
Received 24 May 2013, Accepted 5 January 2014 Published online 17 February 2014 in Wiley Online Library

(wileyonlinelibrary.com)

Targeted comparat time-vary

Romain Neugeb 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

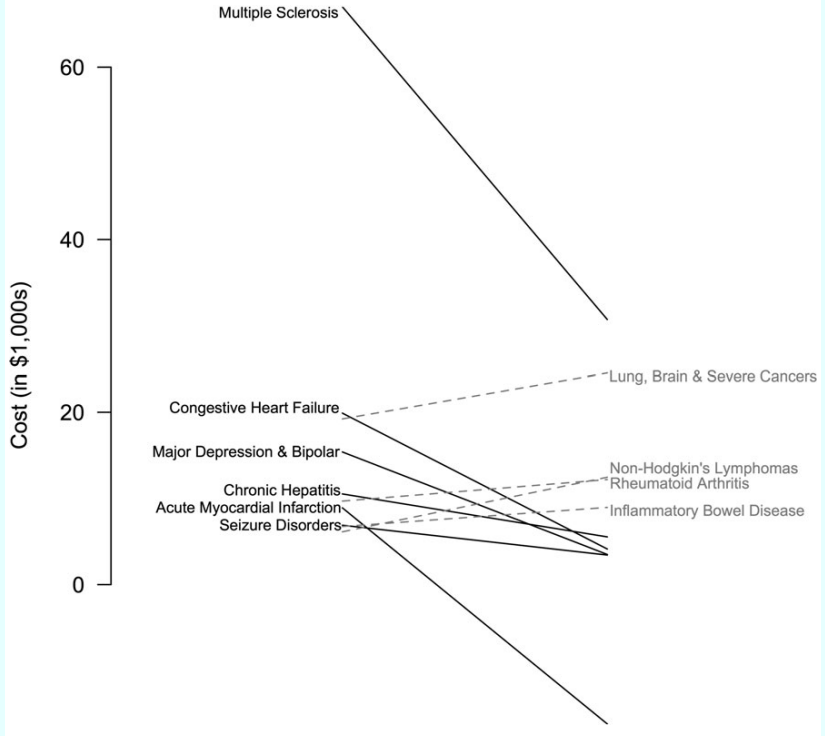
Identifying contributing factors for health care spending

DOI: 10.1111/1475-6773.12848
METHODS ARTICLE

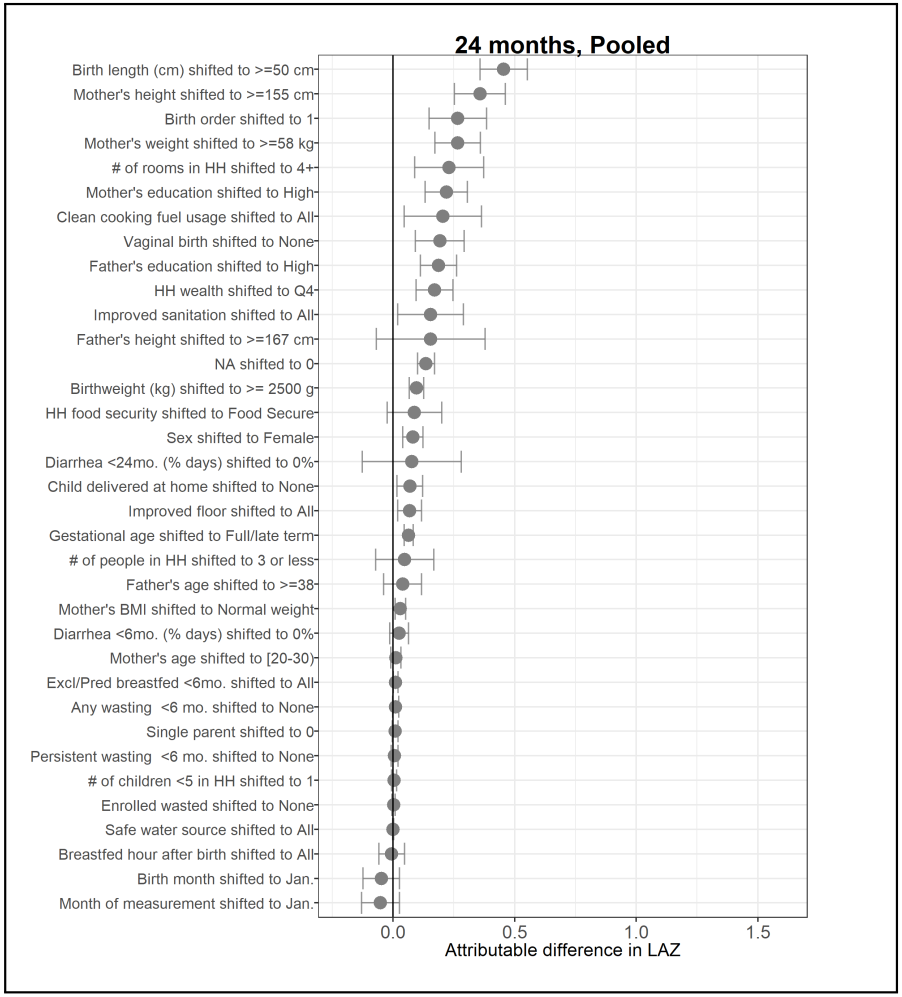
Robust Machine Learning Vari
Importance Analyses of Medic
Conditions for Health Care Spe

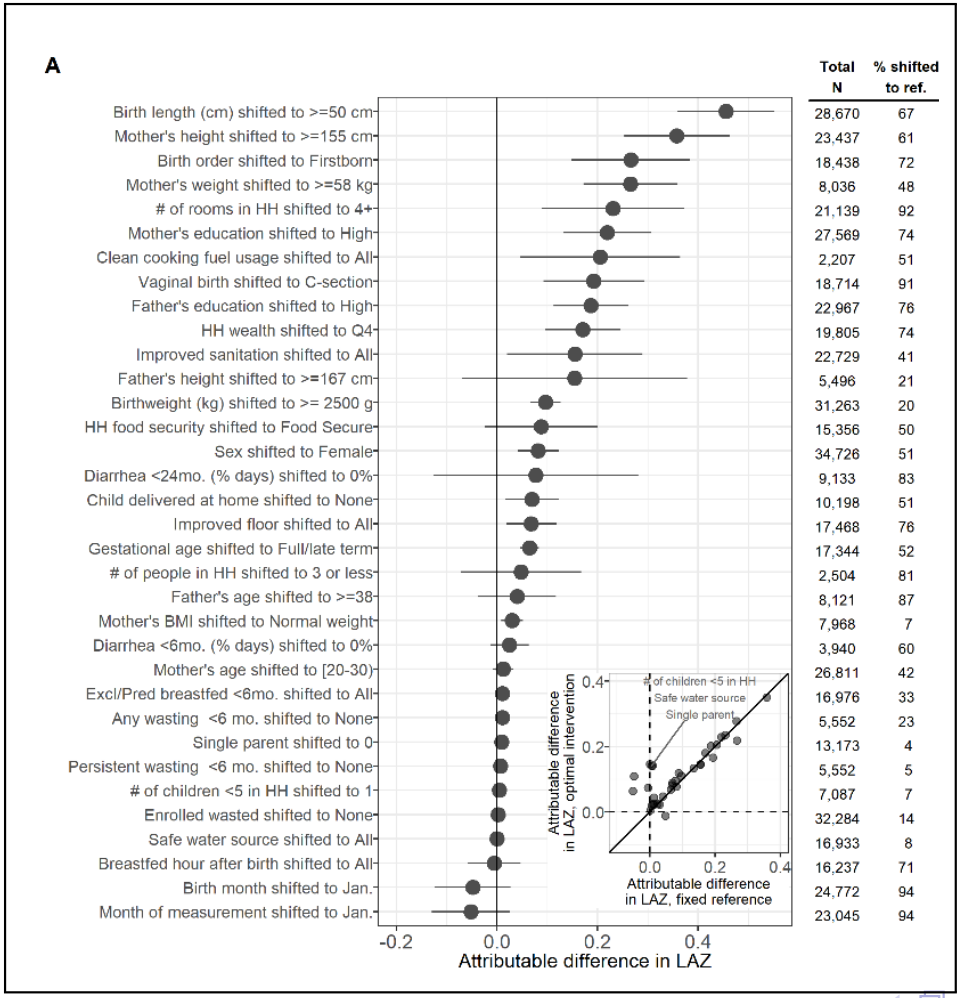
Sherri Rose

Figure 4: Top 10 Largest Targeted Learning Effect Estimates



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





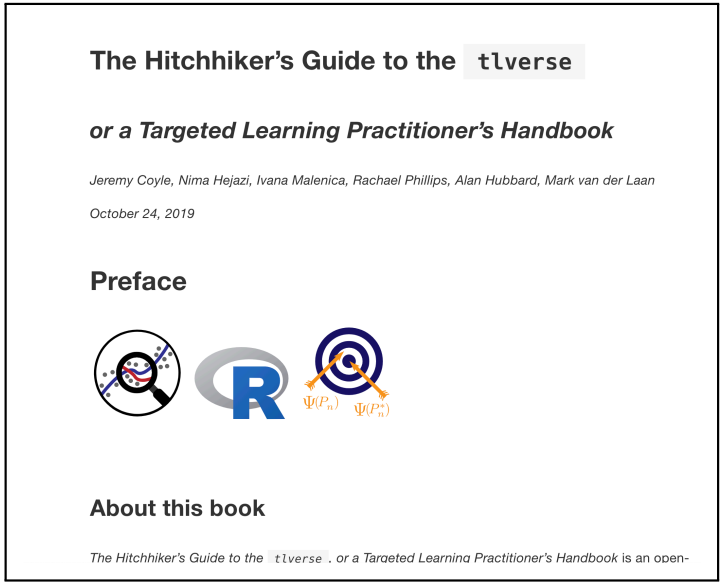
Outline

`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
- Microwave dinners for machine learning

`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

Concluding Remarks

- Targeted Learning learns unbiased and reproducible answers to actionable questions (potentially causal) with confidence, which result in improved policy, treatments, evaluations, etc.
- It integrates **causal inference**, **machine learning**, **statistical theory**.
- **Targeted Learning** *optimally estimates* the (potentially causal) impact of an intervention on an outcome for complex real-world data.
- 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, such as *The Hitchhiker's Guide to the `tlverse`: A Targeted Learning Practitioner's Handbook*.