

**Targeted
Learning (TL)**

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand

Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive
conclusion

Conclusion

Targeted Learning (TL)

The bridge from machine learning
to statistical and causal inference

Professors Alan Hubbard & Mark van der Laan

Division of Biostatistics, University of California, Berkeley

August 7, 2023

American Statistical Association
Joint Statistical Meetings (JSM)

Acknowledgements: Susan Gruber, Ivana Malenica and Rachael Phillips



Traditional toolbox for statistics

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

Real-World
Data Science
with TL

TL Roadmap for Causal and Statistical Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand

Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion

Goal	Type of Data			
	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

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

Real-World
Data Science
with TL

TL Roadmap for Causal and Statistical Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand

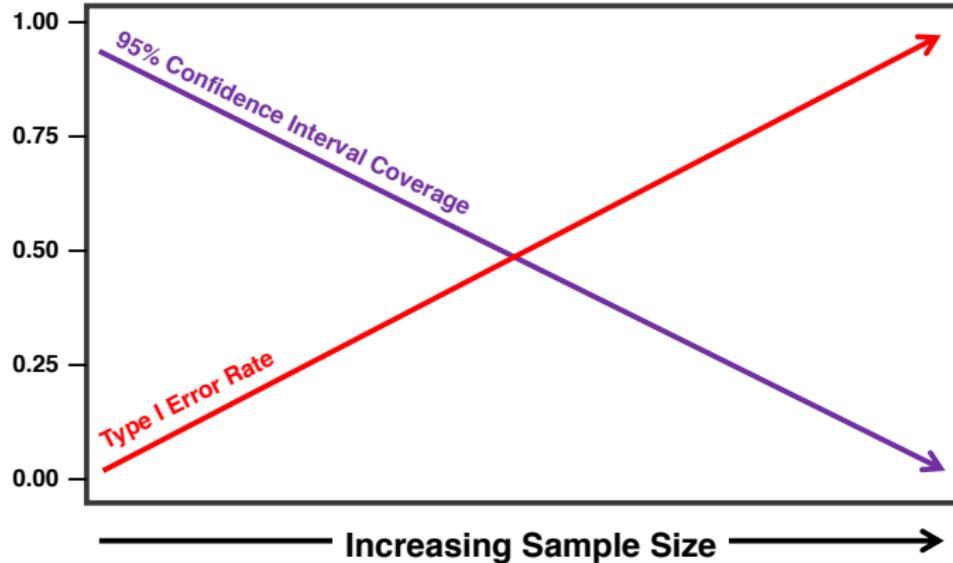
Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion



Post-hoc model manipulation

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

Real-World
Data Science
with TL

TL Roadmap for Causal and Statistical Inference

Describe study
Specify realistic statistical model
Define statistical estimand

Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive conclusion

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.



Why care about statistical inference?

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

Real-World
Data Science
with TL

TL Roadmap for Causal and Statistical Inference

- Describe study
- Specify realistic statistical model
- Define statistical estimand
- Causal estimand
- Statistical estimand
- Construct estimator
- Obtain inference
- Make substantive conclusion

Conclusion

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. Simmons¹, Leif D. Nelson², and Uri Simonsohn¹

¹The Wharton School, University of Pennsylvania, and ²Haas 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



Public health and clinical medicine rely on real-world data (RWD) for insight and evidence

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

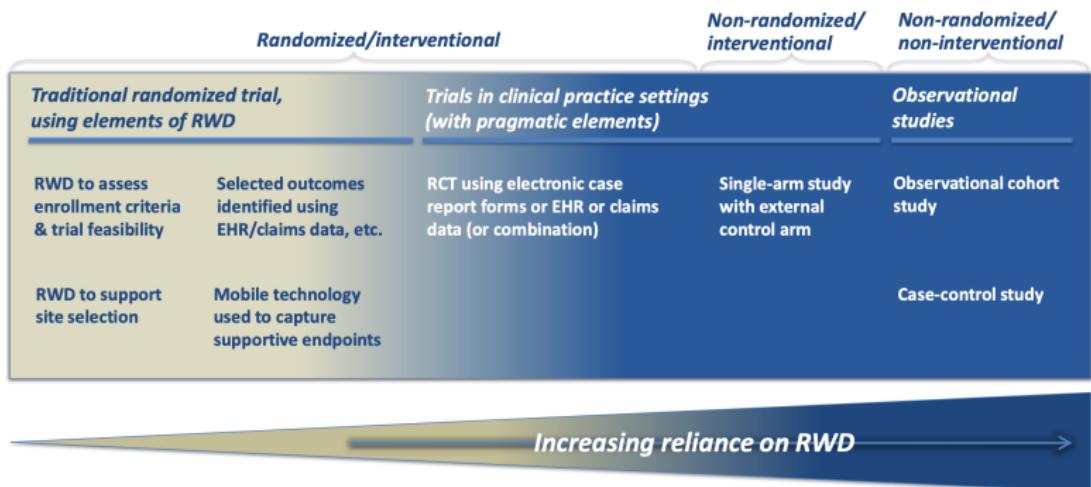
Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand

Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive
conclusion

Conclusion



Courtesy of "FDA Real-World Evidence Program" Webinar by John Concato on 4 August 2021

Statistical challenges in RWD science are addressed by TL

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand

Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion

<i>Randomized/interventional</i>	<i>Non-randomized/ interventional</i>	<i>Non-randomized/ non-interventional</i>
<i>Traditional randomized trial, using elements of RWD</i>	<i>Trials in clinical practice settings (with pragmatic elements)</i>	<i>Observational studies</i>
RWD to assess enrollment criteria & trial feasibility	Selected outcomes identified using EHR/claims data, etc.	RCT using electronic case report forms or EHR or claims data (or combination) Single-arm study with external control arm

RWD Challenges

- Selection bias
- Intercurrent events
- Informative missingness
- Treatment by indication
- High dimensional covariates
- Outcome measurement error
- Statistical model misspecification
- Differences between external controls and single trial arm RCT

*Targeted Learning
path supports regulatory
decision making*

TL provides a systematic process for RWD science and real-world evidence (RWE) evaluation

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

- Describe study
- Specify realistic statistical model
- Define statistical estimand
- Causal estimand
- Statistical estimand
- Construct estimator
- Obtain inference
- Make substantive conclusion

Conclusion

<i>Randomized/interventional</i>	<i>Non-randomized/interventional</i>	<i>Non-randomized/non-interventional</i>
<i>Traditional randomized trial, using elements of RWD</i>	<i>Trials in clinical practice settings (with pragmatic elements)</i>	<i>Observational studies</i>
RWD to assess enrollment criteria & trial feasibility	Selected outcomes identified using EHR/claims data, etc.	RCT using electronic case report forms or EHR or claims data (or combination)
RWD to support site selection	Mobile technology used to capture supportive endpoints	Single-arm study with external control arm
		Observational cohort study
		Case-control study

RWD Challenges

- ❑ Selection bias
- ❑ Intercurrent events
- ❑ Informative missingness
- ❑ Treatment by indication
- ❑ High dimensional covariates
- ❑ Outcome measurement error
- ❑ Statistical model misspecification
- ❑ Differences between external controls and single trial arm RCT

Targeted Learning path supports regulatory decision making

Targeted Learning

- ✓ Roadmap for causal and statistical inference
- ✓ Realistic statistical model
- ✓ Statistical estimand approximates answer to causal question
- ✓ Flexible estimation and dimension reduction with Super Learner
- ✓ Model-free sensitivity analysis
- ✓ Generate RWE with confidence

TL is a subfield of statistics

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand

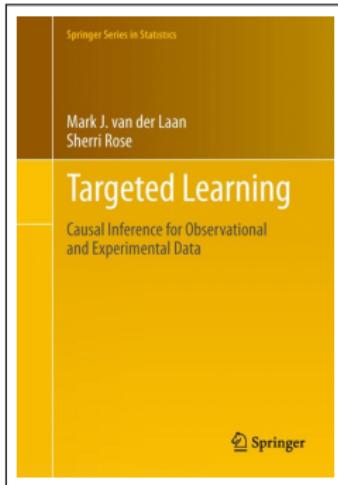
Statistical estimand

Construct estimator

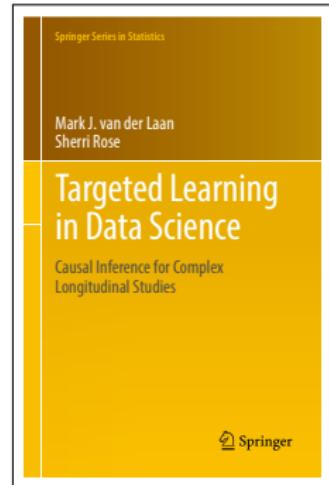
Obtain inference

Make substantive
conclusion

Conclusion



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.

The Hitchhiker's Guide to the tlverse

Applications of TL in the real world

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

Real-World Data Science with TL

TL Roadmap for Causal and Statistical Inference

- Describe study
- Specify realistic statistical model
- Define statistical estimand
- Causal estimand
- Statistical estimand
- Construct estimator
- Obtain inference
- Make substantive conclusion

Conclusion

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. Agyieko, J. Kabami, N. Sang, T. Liegler, G. Chamie, C.S. Camlin, V. Jain, K. Kadefio, M. Atukunda, T. Ruel, S.B. Shade, E. Ssemmondo, D.M. Byonanebye, F. Mwangwa, A. Owaramagise, W. Oollo, D. Black, K. Smyman, 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 Wafulila, 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

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. Moiel, E.L. Moss, P. Njuguna, C.F. Ockenhouse, B. Ragama Oguta, W. Otiemo, L. Otiemo, K. Otiemo, S. Owusu-Agyei, D.J. Park, K. Pelté, 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

© Health Research and Educational Trust

DOI: 10.1111/1475-6773.12848

METHODS ARTICLE

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

Sherri Rose 

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



Better clinical decisions from observational data

Targeted
Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

- Describe study
- Specify realistic statistical model
- Define statistical estimand
- Causal estimand
- Statistical estimand
- Construct estimator
- Obtain inference
- Make substantive conclusion

Conclusion

Statistics
in Medicine

Research Article

Received 24 May 2013,

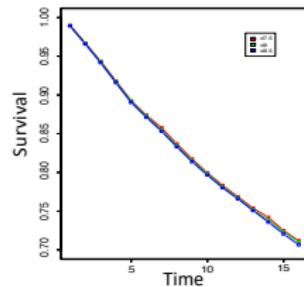
Accepted 5 January 2014

Published online 17 February 2014 in Wiley Online Library

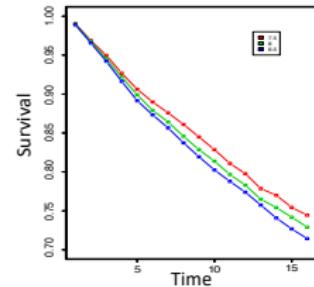
(wileyonlinelibrary.com) DOI: 10.1002/sim.6099

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

Romain Neugebauer,^{a*}[†] Julie A. Schmittield^a and
Mark J. van der Laan^b



Standard methods: No benefit to more aggressive intensification strategy



Targeted Learning: More aggressive intensification protocols result in better outcomes

TL for Optimal Treatment Variable Importance in Meta-analysis of Childhood Development Studies

Targeted
Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study

Specify realistic
statistical model

Define statistical
estimand

Causal estimand

Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion

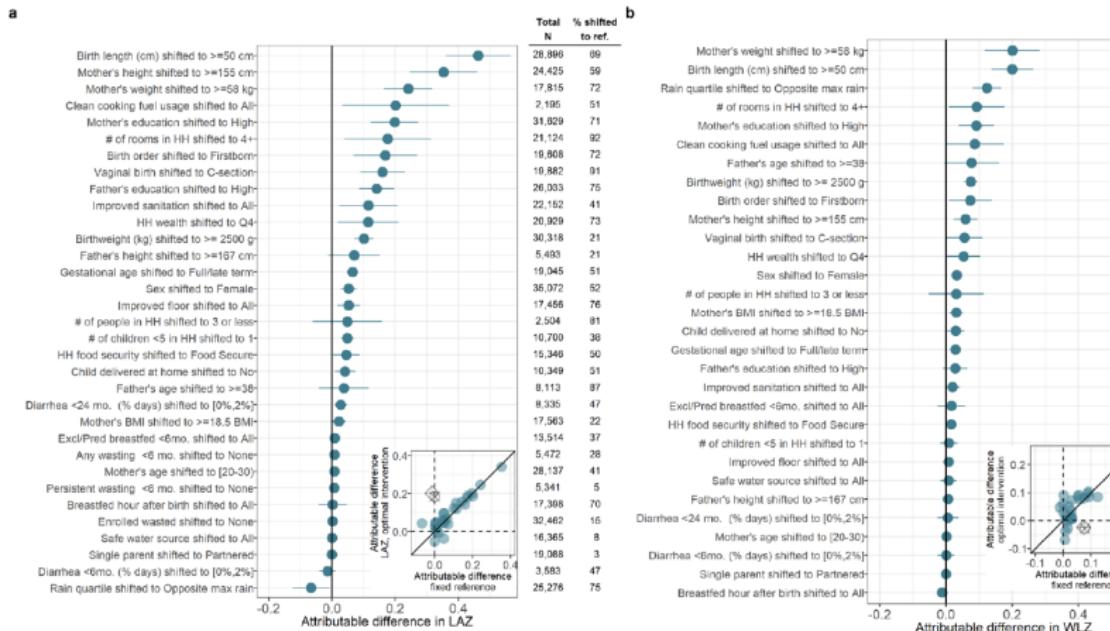


Figure 2 | Rank-ordered attributable differences between child, parental, and household characteristics and population attributable differences in anthropometry Z-scores.

Evaluating counterfactual HIV-1 infection risk after shifting vaccine-induced immune response markers

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand
Statistical estimand

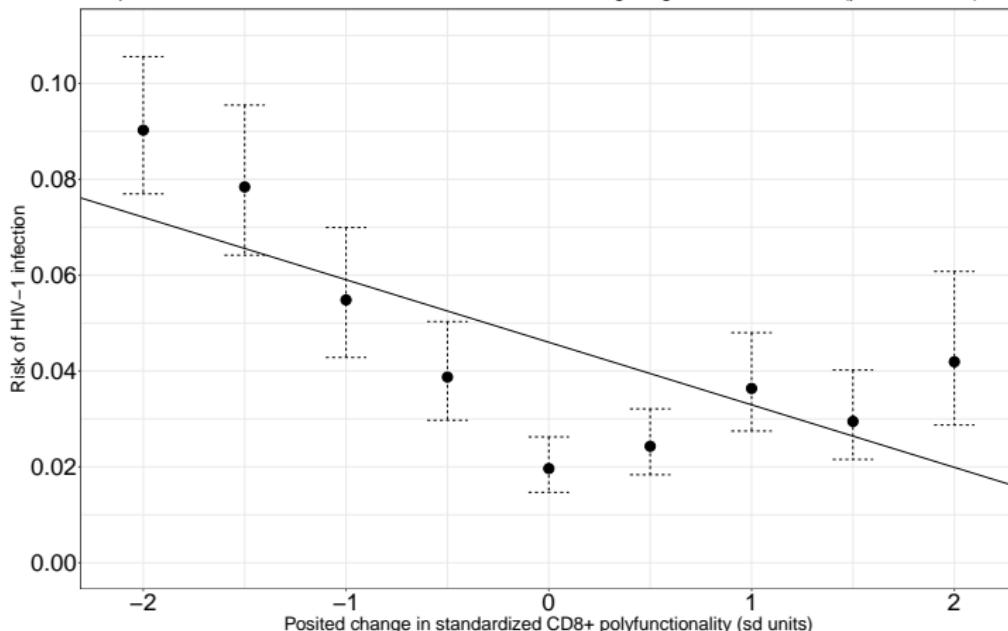
Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion

TML estimates of mean counterfactual HIV-1 infection risk under shifted CD8+ polyfunctionality
with pointwise confidence intervals and summarization via working marginal structural model ($\hat{\beta}_{TMLE} = -0.013$)



Hejazi, van der Laan, Janes, Gilbert, Benkeser "Efficient nonparametric inference
on the effects of stochastic interventions under two-phase sampling, with
applications to vaccine efficacy trials" *Biometrics* (2021). 10.1111/biom.13375



Estimate impacts on COVID-19 infection rates by policy shift interventions

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

Real-World Data Science with TL

TL Roadmap for Causal and Statistical Inference

Describe study
Specify realistic statistical model
Define statistical estimand

Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive conclusion

Conclusion



Nugent and Balzer. "A demonstration of Modified Treatment Policies to evaluate shifts in mobility and COVID-19 case rates in US counties." AJE (2023).

TL approximates answers to statistical and causal questions with confidence intervals

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

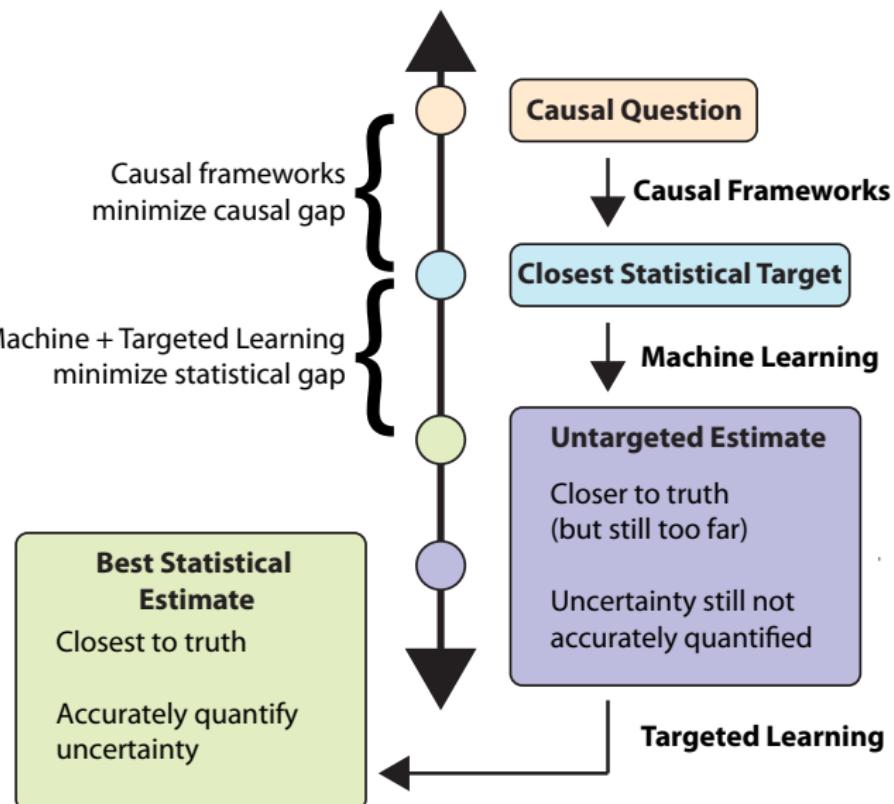
Human Art in Statistics

Real-World Data Science with TL

TL Roadmap for Causal and Statistical Inference

- Describe study
- Specify realistic statistical model
- Define statistical estimand
- Causal estimand
- Statistical estimand
- Construct estimator
- Obtain inference
- Make substantive conclusion

Conclusion



The roadmap for learning from data

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

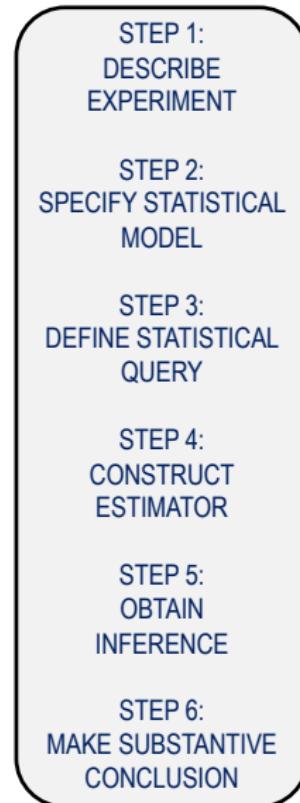
Human Art in Statistics

Real-World
Data Science
with TL

TL Roadmap for Causal and Statistical Inference

- Describe study
- Specify realistic statistical model
- Define statistical estimand
- Causal estimand
- Statistical estimand
- Construct estimator
- Obtain inference
- Make substantive conclusion

Conclusion



Step 1: What is the experiment that generated the data?

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

Real-World Data Science with TL

TL Roadmap for Causal and Statistical Inference

Describe study

Specify realistic
statistical model

Define statistical
estimand

Causal estimand

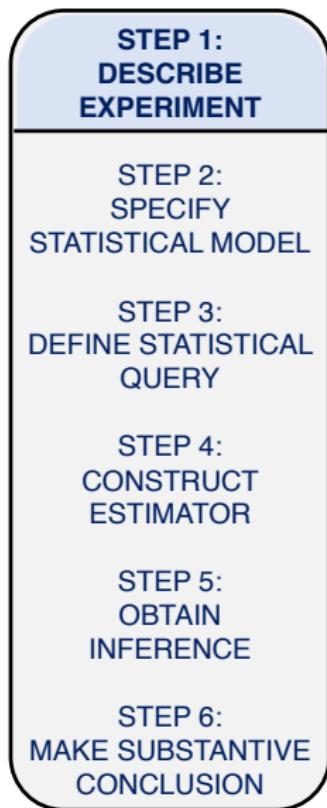
Statistical estimand

Construct estimator

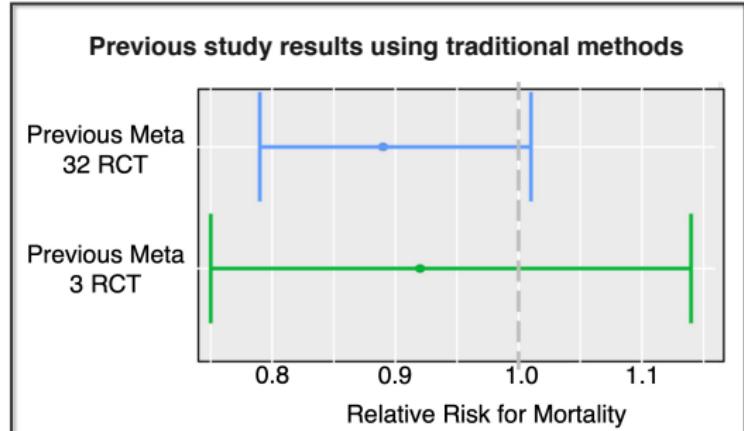
Obtain inference

Make substantive
conclusion

Conclusion



*Three multi-national RCTs assessing
impact of corticosteroids on mortality
among septic shock patients*



Target population of interest

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

Real-World
Data Science
with TL

TL Roadmap for Causal and Statistical Inference

Describe study

Specify realistic
statistical model

Define statistical
estimand

Causal estimand

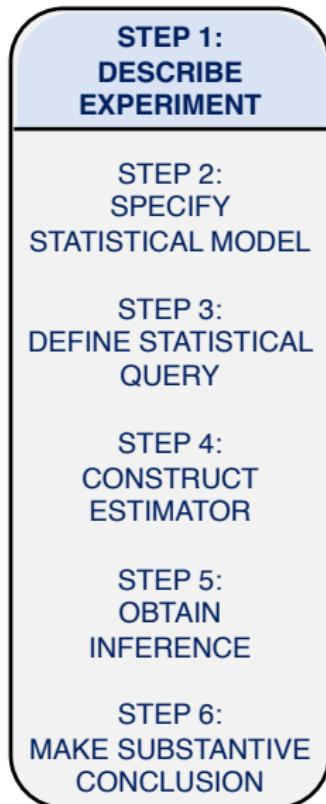
Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion



***Three multi-national RCTs assessing
impact of corticosteroids on mortality
among septic shock patients***

Pooled sample of $n = 1,300$ adults in septic shock

Observed data structure

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

Real-World
Data Science
with TL

TL Roadmap for Causal and Statistical Inference

Describe study

Specify realistic
statistical model

Define statistical
estimand

Causal estimand

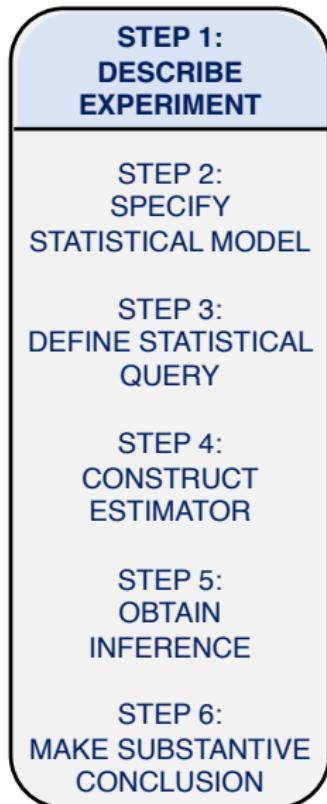
Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion



***Three multi-national RCTs assessing
impact of corticosteroids on mortality
among septic shock patients***

Pooled sample of $n = 1,300$ adults in septic shock

BMI	age	sex	steroid treatment	1-month mortality
21	65	F	1	1
22.3	28	F	0	0
19.4	49	F	1	0
24	77	M	1	0

Directed Acyclic Graph (DAG)

Targeted Learning (TL)

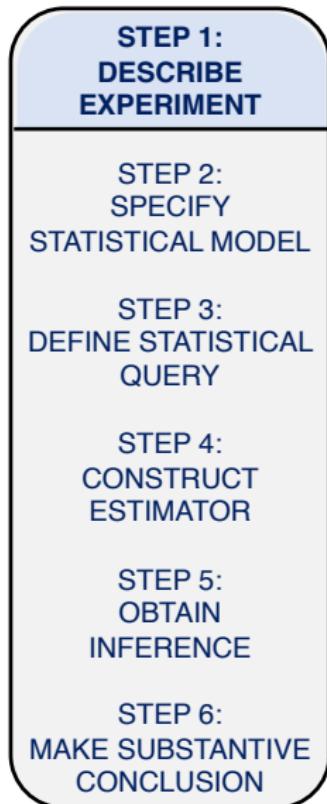
Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

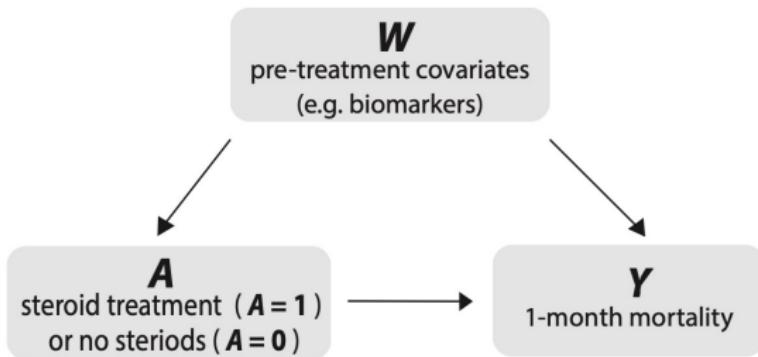
TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand
Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive
conclusion
Conclusion

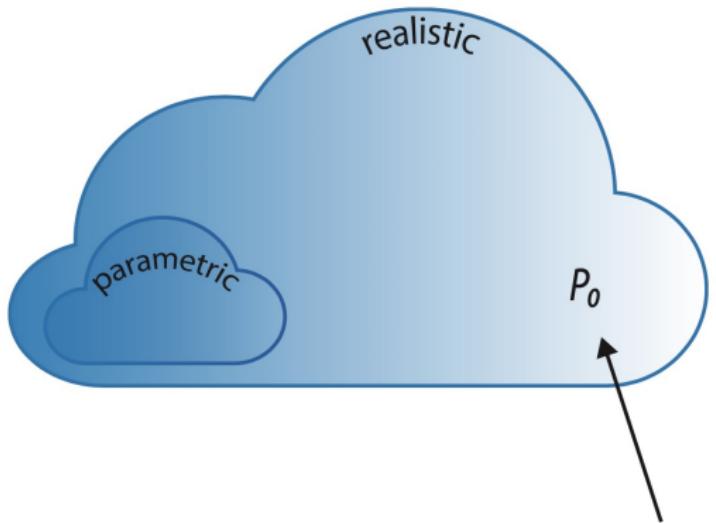
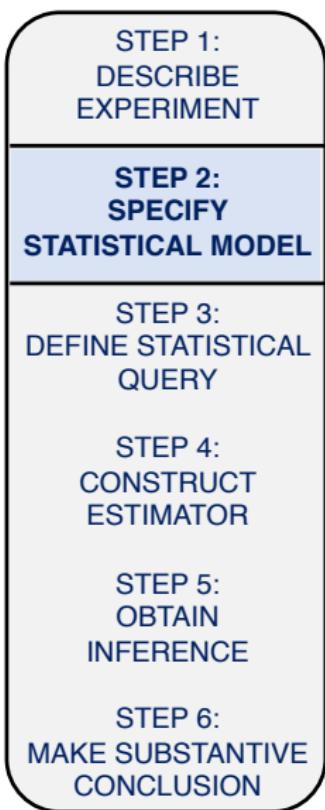


*Three multi-national RCTs assessing
impact of corticosteroids on mortality
among septic shock patients*

Pooled sample of $n = 1,300$ adults in septic shock



Step 2: What is known about stochastic relations of the observed variables?



True data-generating process (DGP)

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

Real-World Data Science with TL

TL Roadmap for Causal and Statistical Inference

Describe study

Specify realistic statistical model

Define statistical estimand

Causal estimand

Statistical estimand

Construct estimator

Obtain inference

Make substantive conclusion

Conclusion

What happens when the statistical model is misspecified (i.e. does not contain P_0 , the DGP)?

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

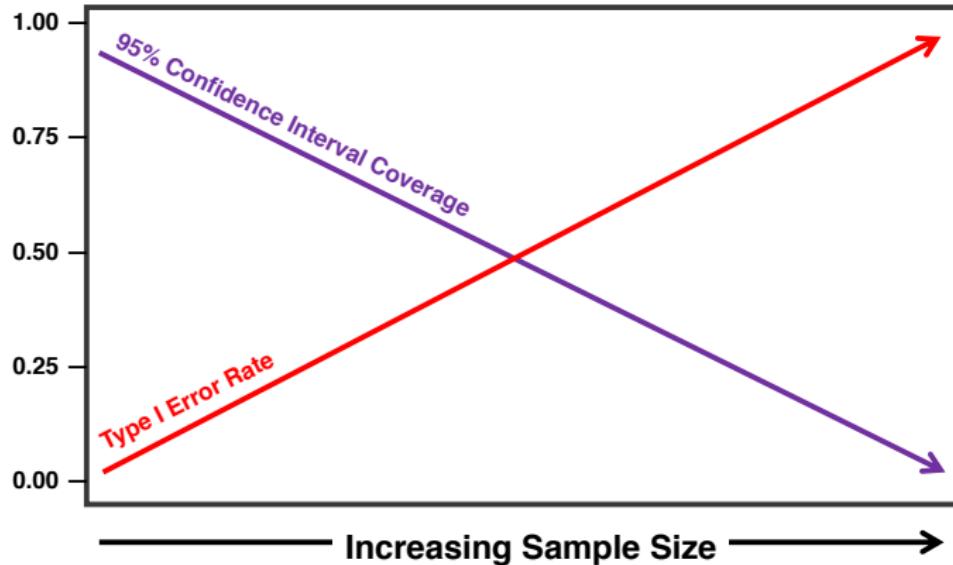
Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand

Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive
conclusion

Conclusion



Step 3a: What is the target causal estimand that we aim to identify from the data?

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

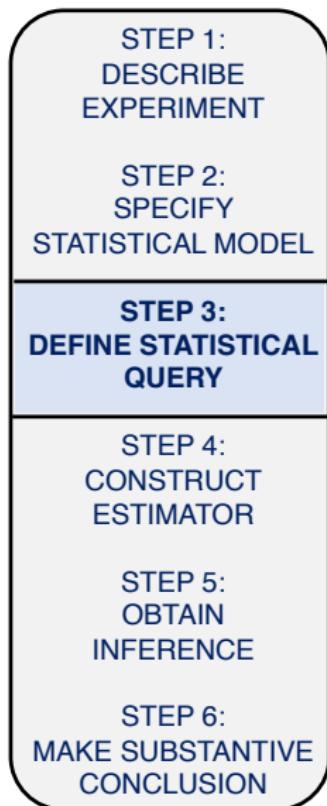
Real-World
Data Science
with TL

TL Roadmap for Causal and Statistical Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand

Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive
conclusion

Conclusion



What is the causal risk ratio in mortality between treatment groups?

$$\psi_{causal} = \frac{P(Y_1 = 1)}{P(Y_0 = 1)}$$

Proportion of subjects in the population of interest that would have died had they all received steroids

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

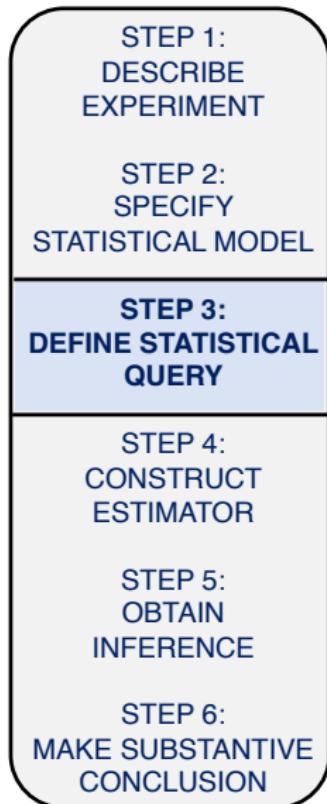
Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand

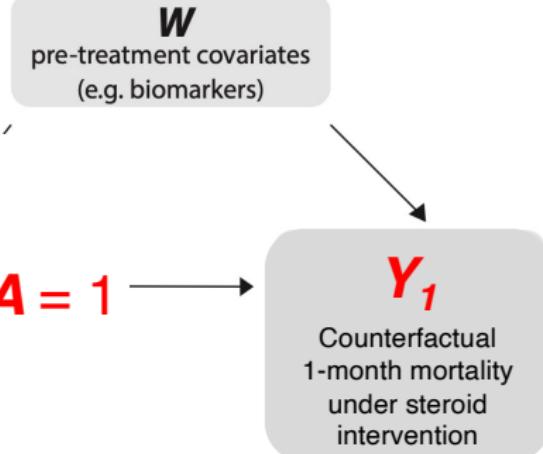
Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive
conclusion

Conclusion



What is the causal risk ratio in mortality between treatment groups?

$$\psi_{causal} = \frac{P(Y_1 = 1)}{P(Y_0 = 1)}$$



Proportion of subjects in the target population that would have died had they all not received steroids

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

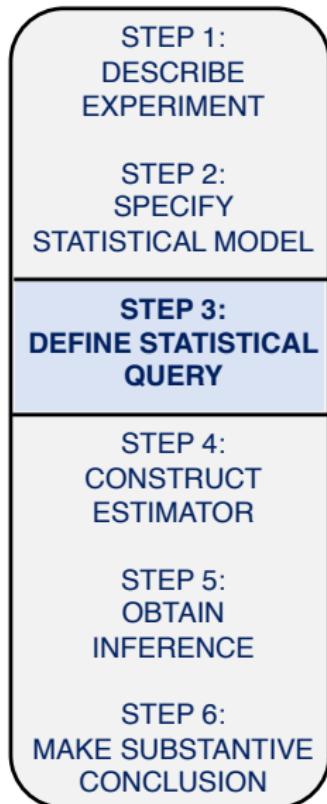
Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand

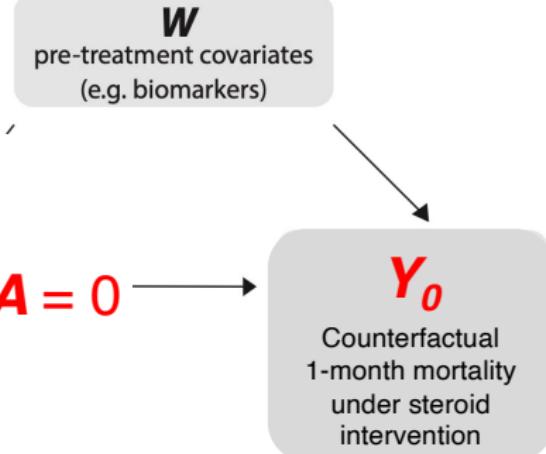
Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive
conclusion

Conclusion



What is the causal risk ratio in mortality between treatment groups?

$$\psi_{causal} = \frac{P(Y_1 = 1)}{P(Y_0 = 1)}$$



Causal target parameters are functions of the full data under the intervention(s) of interest

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

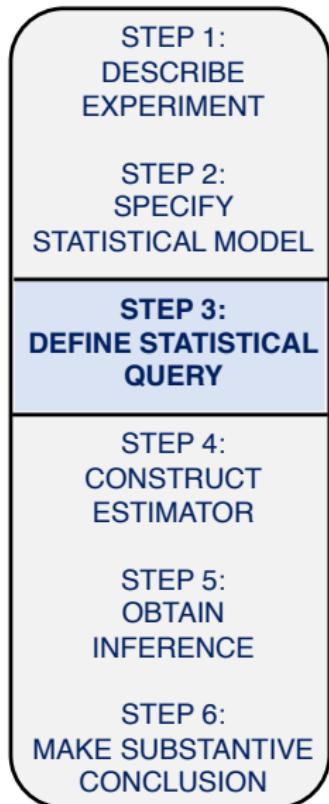
Causal estimand
Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion



What is the causal risk ratio in mortality between treatment groups?

$$\psi_{causal} = \frac{P(Y_1 = 1)}{P(Y_0 = 1)}$$

W_1	W_2	W_3	A	Y	Y_1	Y_0
21	65	F	1	1	1	?
22.3	28	F	0	0	?	0
19.4	49	F	1	0	0	?
24	77	M	1	0	0	?

Causal identifiability assumptions must hold in order to interpret the estimate causally

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

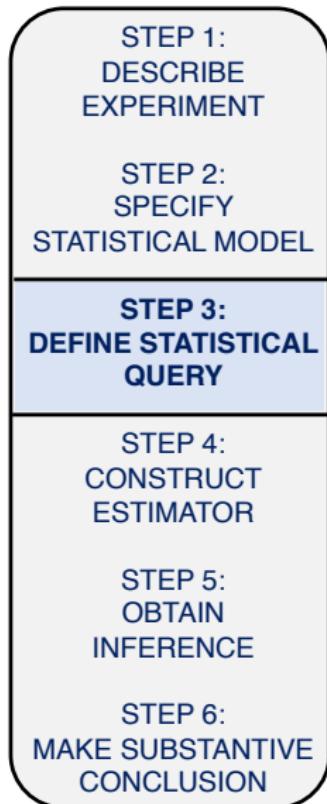
Real-World
Data Science
with TL

TL Roadmap for Causal and Statistical Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand

Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive
conclusion

Conclusion



What is the causal risk ratio in mortality between treatment groups?

$$\psi_{causal} = \frac{P(Y_1 = 1)}{P(Y_0 = 1)}$$

What's needed to *identify* ψ_{causal} from the observed data?

1. No unmeasured confounding / randomization / exchangeability

Some causal identifiability assumptions are also necessary for well-defined statistical estimation

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

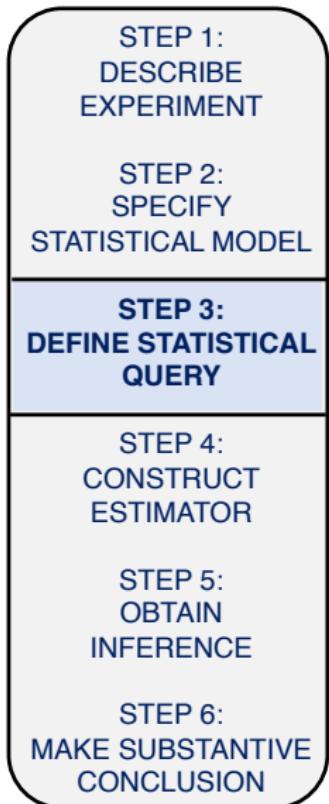
Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand

Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive
conclusion



What is the causal risk ratio in mortality between treatment groups?

$$\psi_{causal} = \frac{P(Y_1 = 1)}{P(Y_0 = 1)}$$

What's needed to ***identify*** ψ_{causal} from the observed data?

1. No unmeasured confounding / randomization / exchangeability
2. Positivity / experimental treatment assignment (ETA) assumption

Conclusion

G-computation identification formula

Targeted
Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand

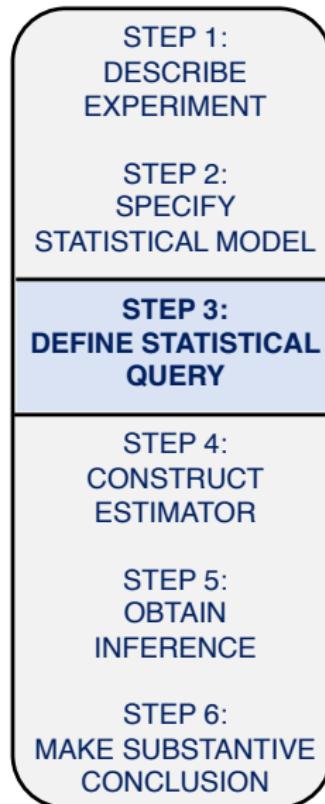
Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion

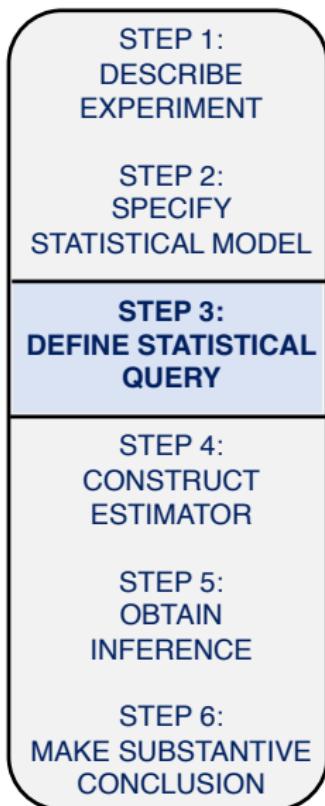


$$\psi_{causal} = \frac{P(Y_1 = 1)}{P(Y_0 = 1)}$$

$$= \frac{\sum_w P(Y = 1|A = 1, W = w)P(W = w)}{\sum_w P(Y = 1|A = 0, W = w)P(W = w)}$$

$$= \psi_{stat}$$

Step 3b: What is the target statistical estimand that we will learn from the data?



What is the risk of mortality between treatment groups when adjusting for covariates?

$$\psi_{stat} =$$

$$\frac{\sum_w P(Y = 1|A = 1, W = w)P(W = w)}{\sum_w P(Y = 1|A = 0, W = w)P(W = w)}$$

Targeted
Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand

Causal estimand

Statistical estimand

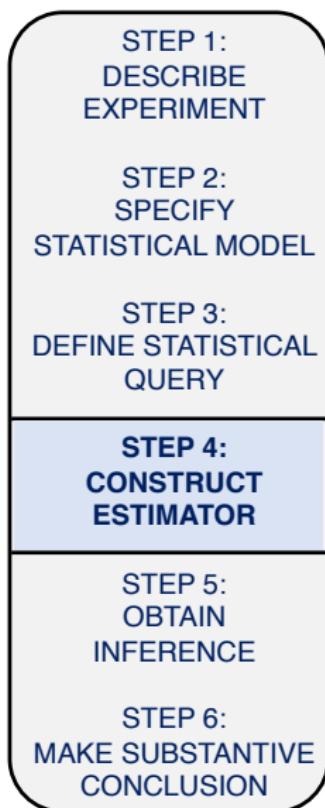
Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion

Step 4: How should we estimate the target estimand?



Statistical properties to consider

- Substitution / plug-in
- Valid inference
- Efficiency
- Ability to optimize finite sample performance

Targeted
Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand
Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion

Targeted Maximum Likelihood Estimation (TMLE)

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

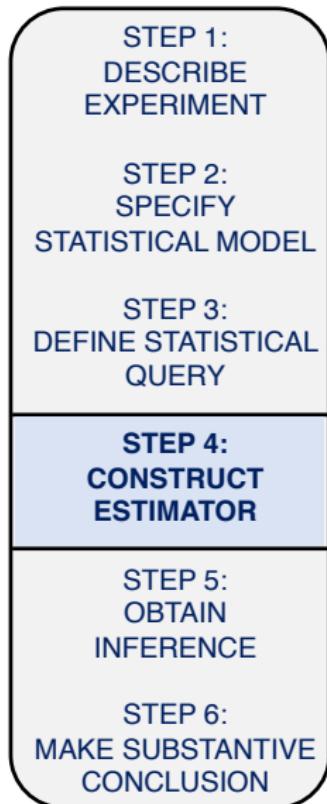
Real-World
Data Science
with TL

TL Roadmap for Causal and Statistical Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand

Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive
conclusion

Conclusion



TMLE

- 1 Initial estimation of $P[Y = 1|A, W]$ with super (machine) learning
- 2 Updating initial estimate to achieve optimal bias-variance trade-off for ψ_{stat}

TMLE estimates are optimal:
plug-in, efficient, unbiased, finite sample robust

TMLE Step 1: Super learner

Targeted
Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand

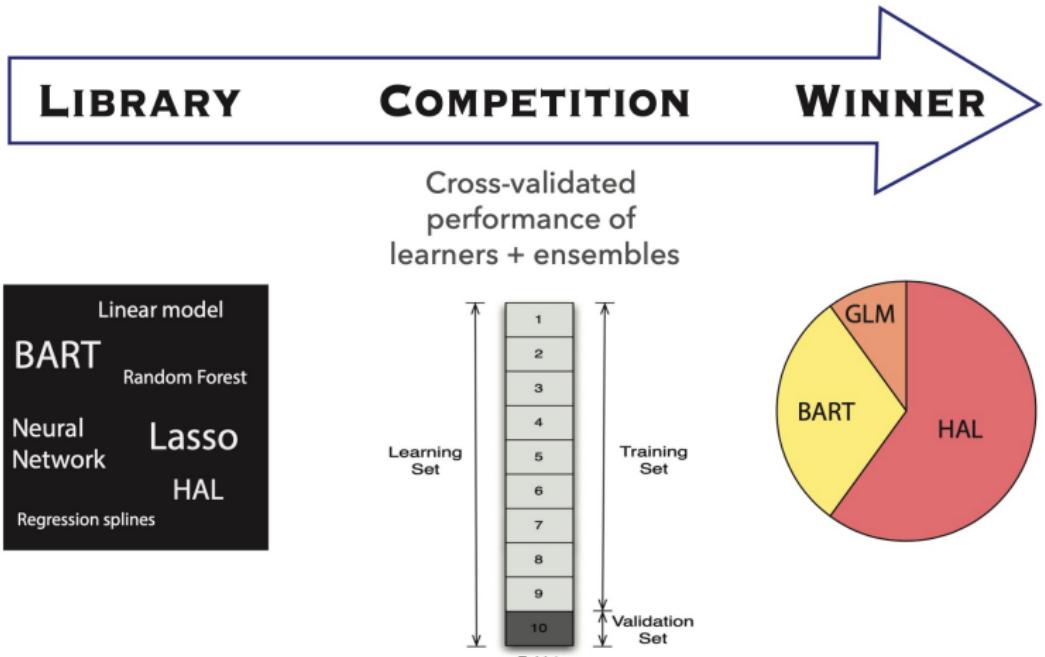
Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion



Hugely advantageous when coupled with NLP-derived covariates with EHR

Tip: Include the Highly Adaptive Lasso (HAL) as a candidate in the SL library

- First estimator that guarantees asymptotically efficient estimation of any pathwise differentiable estimand¹ (e.g., the average causal effect or treatment-specific survival), without enforcing strong smoothness conditions.
- Exceedingly mild assumptions, expected to hold in almost every practical application.
- Can be implemented with standard Lasso software.
- Converges to true function at rate $n^{-1/3}(\log n)^{d/2}$.
- Accommodates a variety of function space specifications.

¹An estimand that is a weakly differentiable functional of the density of the data, the case for most causal inference estimands under positivity.



Illustration in Low Dimensions

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand

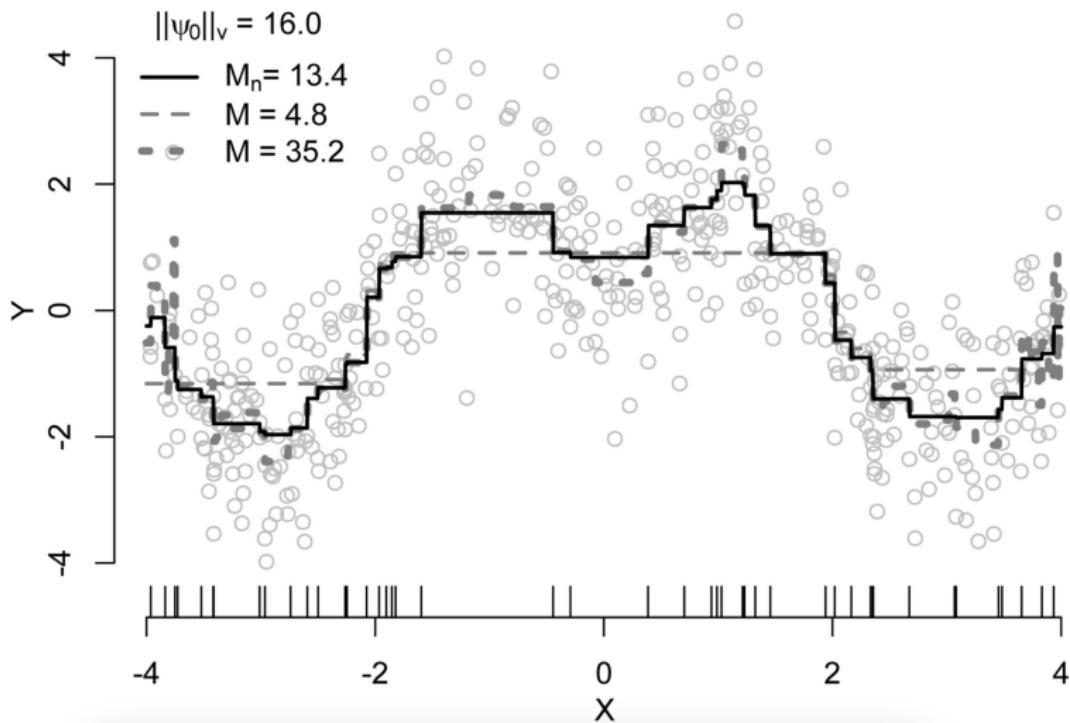
Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion



TMLE Step 2: Targeting follows a path of maximal change in target estimand per unit likelihood

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand
Statistical estimand

Construct estimator

Obtain inference
Make substantive
conclusion

Conclusion

How should we approximate the sampling distribution of our estimator?

Targeted Learning (TL)

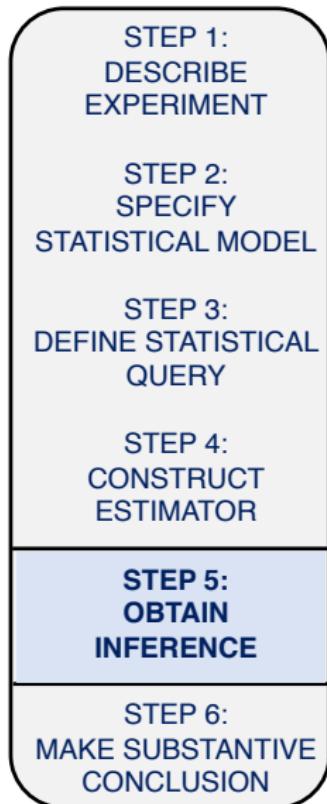
Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

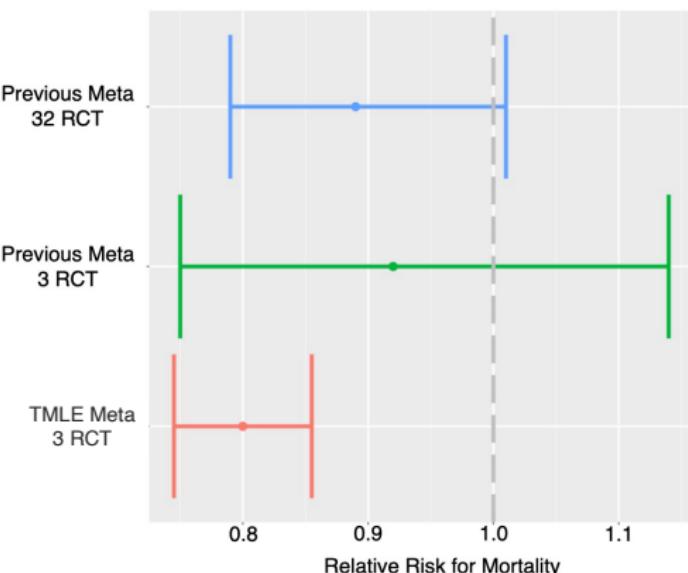
Real-World
Data Science
with TL

TL Roadmap for Causal and Statistical Inference

- Describe study
- Specify realistic statistical model
- Define statistical estimand
- Causal estimand
- Statistical estimand
- Construct estimator
- Obtain inference
- Make substantive conclusion
- Conclusion



Due to targeting (step ②), the TMLE behaves as the *sample mean* of efficient influence function



Arriving at the substantive conclusion

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

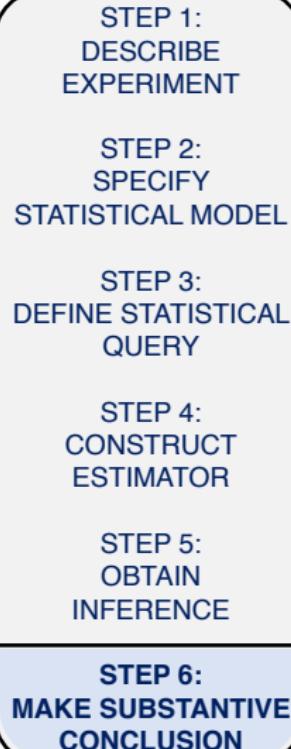
Real-World Data Science with TL

TL Roadmap for Causal and Statistical Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand

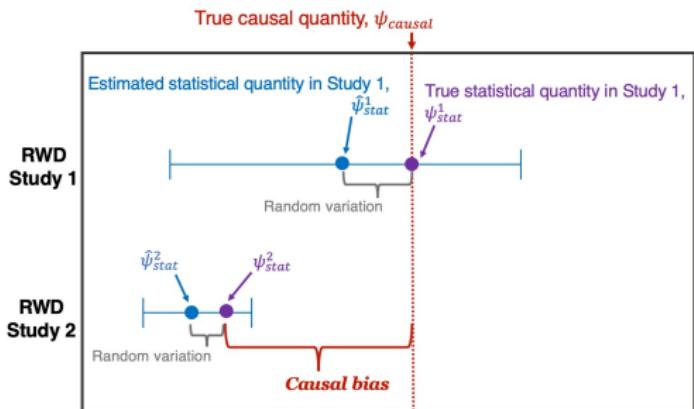
Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive
conclusion

Conclusion



Investigate causal bias with sensitivity analysis

Causal bias: Gap between estimate and truth due to violations of any of the causal assumptions (e.g., unmeasured confounding)*



Sensitivity Analysis: Model-free assessment of how reasonable departures from causal assumptions would impact study findings

* Sensitivity analysis can be extended to incorporate statistical bias

TL-based non-parametric sensitivity analysis: RCT with 25% LTFU example

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal
and
Statistical
Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand

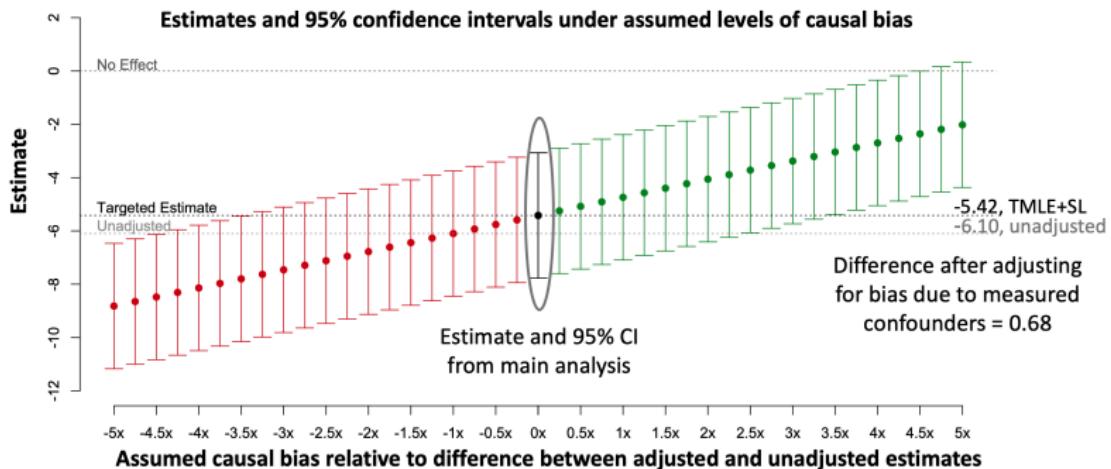
Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion



Courtesy of "Targeted-Learning Based Statistical Analysis Plan" Webinar by Susan Gruber on 28 April 2021



TL-based non-parametric sensitivity analysis: Safety analysis example

Targeted
Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in
Statistics

Real-World
Data Science
with TL

TL Roadmap
for Causal and
Statistical
Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand

Statistical estimand

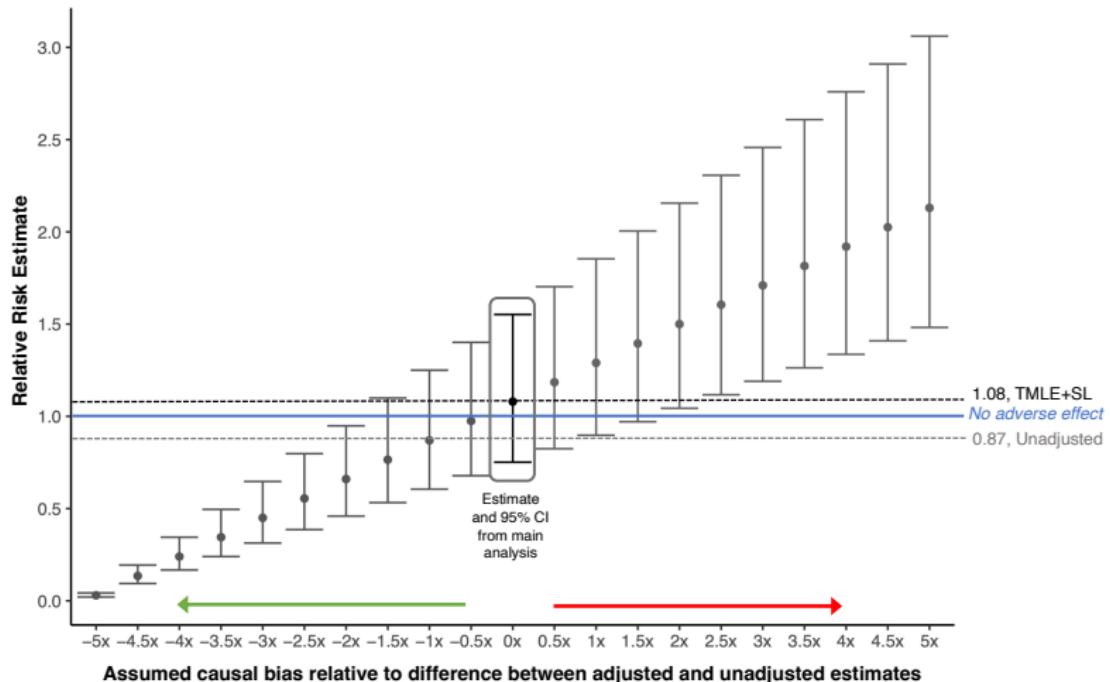
Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion

Relative risk estimates and 95% confidence intervals under assumed levels of causal bias



Possibility to refine question of interest and inform future studies

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

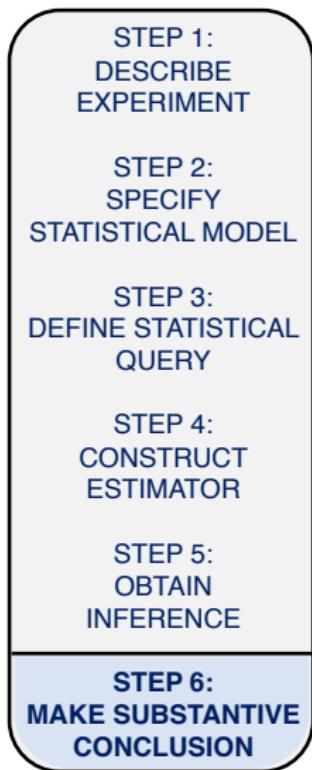
Human Art in Statistics

Real-World Data Science with TL

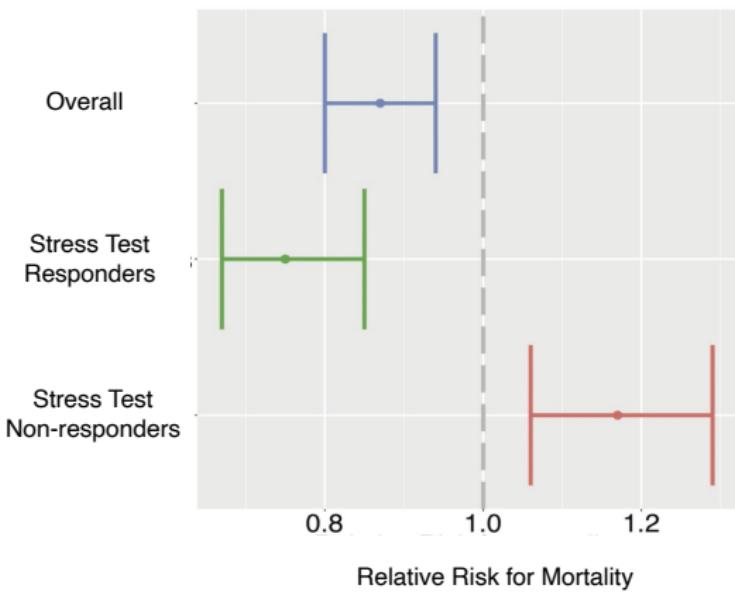
TL Roadmap for Causal and Statistical Inference

Describe study
Specify realistic
statistical model
Define statistical
estimand

Causal estimand
Statistical estimand
Construct estimator
Obtain inference
Make substantive
conclusion



What subgroup of patients in septic shock benefit from corticosteroids?



Active TL collaborations and software

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Human Art in Statistics

Real-World Data Science with TL

TL Roadmap for Causal and Statistical Inference

Describe study
Specify realistic
statistical model

Define statistical
estimand

Causal estimand

Statistical estimand

Construct estimator

Obtain inference

Make substantive
conclusion

Conclusion

- The Center for Targeted Machine Learning (CTML) at UC Berkeley: <https://ctml.berkeley.edu>
- Government: US Food and Drug Administration (FDA) and National Institute of Allergy and Infectious Diseases (NIAID), California Department of Public Health
- Private sector: Genentech, Gilead, Kaiser Permanente, Netflix, Pandora, Accenture, Novo Nordisk
- Nonprofit: Patient-Centered Outcomes Research Institute (pcori), Bill and Melinda Gates Foundation
- Software:
 - The `tlverse` software ecosystem for TL in R:
<https://tlverse.org>
 - Additional TL software: `tmle`, `ltmle`, `SuperLearner`

Thank you!

Targeted Learning (TL)

Professors
Alan Hubbard
& Mark van
der Laan

Feel free to email us to ask questions, request learning resources, get involved in CTML and/or TL research, etc.

Alan's email: hubbard@berkeley.edu

Mark's email: laan@berkeley.edu

Nima's email: nhejazi@hsph.harvard.edu

Ivana's email: imalenica@fas.harvard.edu

Rachael's email: rachaelvphillips@berkeley.edu

Human Art in Statistics

Real-World Data Science with TL

TL Roadmap for Causal and Statistical Inference

Describe study
Specify realistic statistical model

Define statistical estimand

Causal estimand

Statistical estimand

Construct estimator

Obtain inference

Make substantive conclusion

Conclusion