Targeted Learning

The bridge from machine learning to statistical and causal inference

Mark van der Laan

Jiann-Ping Hsu/Karl E. Peace Professor in Biostatistics & Statistics
University of California, Berkeley

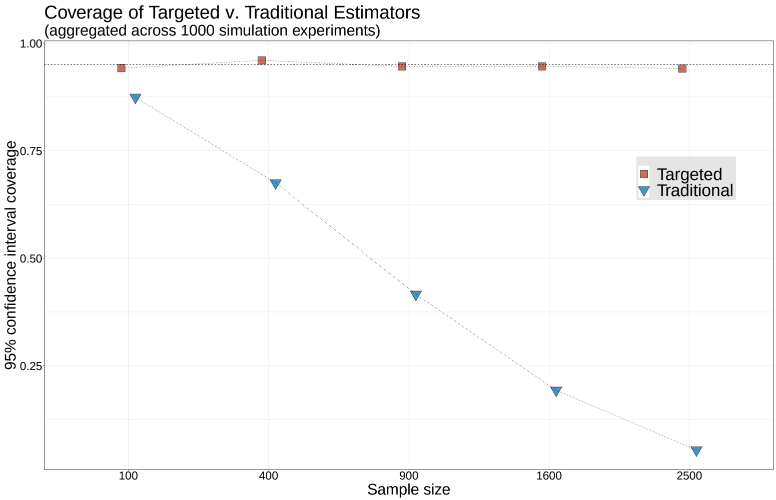
4 March 2020
Alan Turing Institute

Acknowledgements: Rachael Phillips, Ivana Malenica,
Chris Kennedy, Aurelien Bibaut, Nima Hejazi and Jonathan Levi

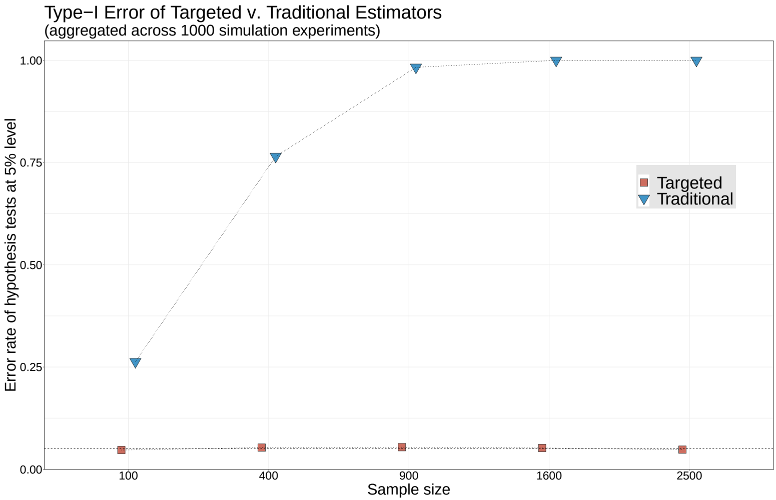
Traditional toolbox for statistics

	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 ttest	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*

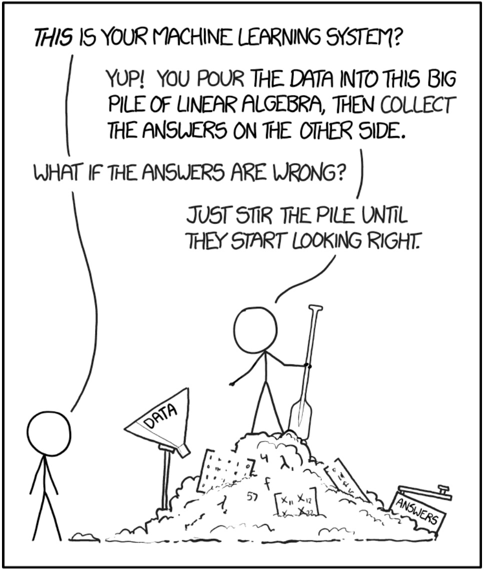
Performance of traditional tools



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]

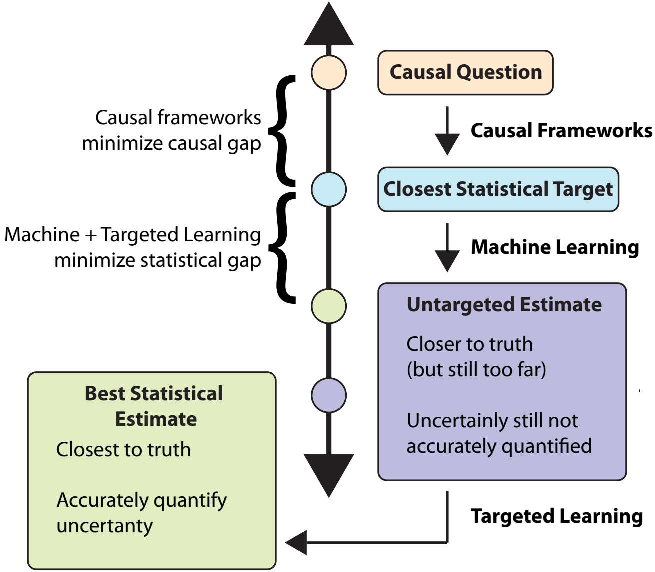
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

The Statistical Crisis in Science

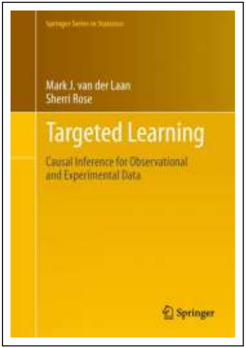
Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Andrew Gelman and Eric Loken

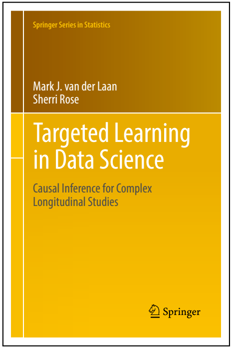
Targeted Learning for answering statistical and causal questions with confidence intervals



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.

The Hitchhiker's Guide to the `tlverse`

Real-world applications of Targeted Learning

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

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

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

Statistics in Medicine

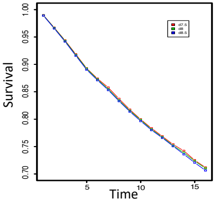
Research Article

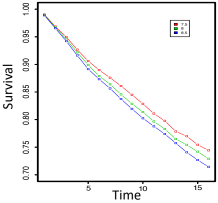
Received 24 May 2013, Accepted 5 January 2014 Published online 17 February 2014 in Wiley Online Library

(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]





Standard methods: No benefit to more aggressive intensification strategy

Targeted Learning: More aggressive intensification protocols result in better outcomes

The roadmap for statistical learning

STEP 1:
DESCRIBE
EXPERIMENT

STEP 2:
SPECIFY
STATISTICAL MODEL

STEP 3:
DEFINE
STATISTICAL QUERY

STEP 4:
CONSTRUCT
ESTIMATOR

STEP 5:
OBTAIN
INFERENCE

What is the experiment that generated the data?

STEP 1: DESCRIBE EXPERIMENT

STEP 2: SPECIFY STATISTICAL MODEL

STEP 3: DEFINE STATISTICAL QUERY

STEP 4: CONSTRUCT ESTIMATOR

STEP 5: OBTAIN INFERENCE

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

What is the experiment that generated the data?

STEP 1: DESCRIBE EXPERIMENT

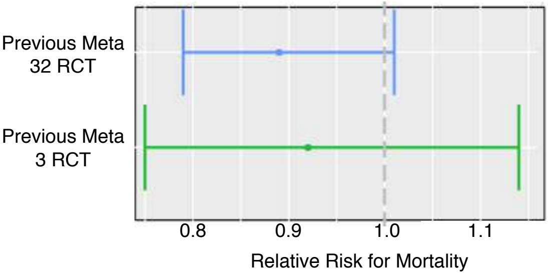
STEP 2: SPECIFY STATISTICAL MODEL

STEP 3: DEFINE STATISTICAL QUERY

STEP 4: CONSTRUCT ESTIMATOR

STEP 5: OBTAIN INFERENCE

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



What is the experiment that generated the data?

STEP 1: DESCRIBE EXPERIMENT

STEP 2: SPECIFY STATISTICAL MODEL

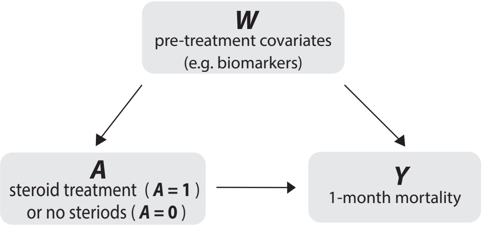
STEP 3: DEFINE STATISTICAL QUERY

STEP 4: CONSTRUCT ESTIMATOR

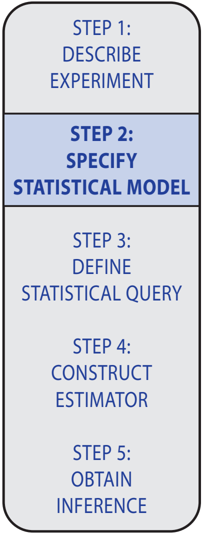
STEP 5: OBTAIN INFERENCE

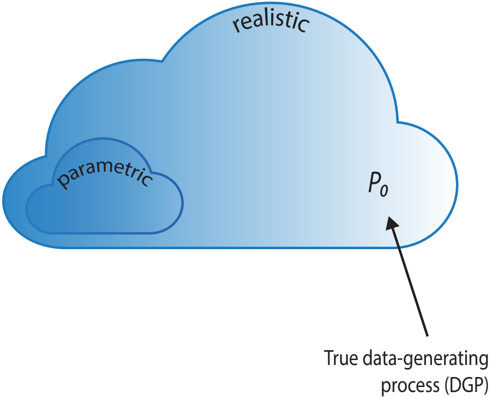
Three multi-national RCTs assessing impact of corticosteroids on mortality among septic shock patients

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

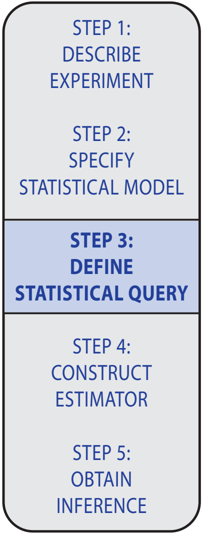


What is known about stochastic relations for the observed variables?





What is the target estimand that we want to learn from the data?



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

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

How should we estimate the target estimand?

STEP 1: DESCRIBE EXPERIMENT

STEP 2: SPECIFY STATISTICAL MODEL

STEP 3: DEFINE STATISTICAL QUERY

STEP 4: CONSTRUCT ESTIMATOR

STEP 5: OBTAIN INFERENCE

TARGETED MAXIMUM LIKELIHOOD ESTIMATION

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

TMLE estimates are optimal:

plug-in, efficient, unbiased, finite sample robust

How should we approximate the sampling distribution of our estimator?

STEP 1:
DESCRIBE
EXPERIMENT

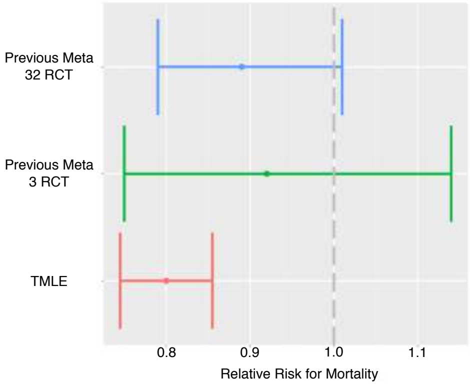
STEP 2:
SPECIFY
STATISTICAL MODEL

STEP 3:
DEFINE
STATISTICAL QUERY

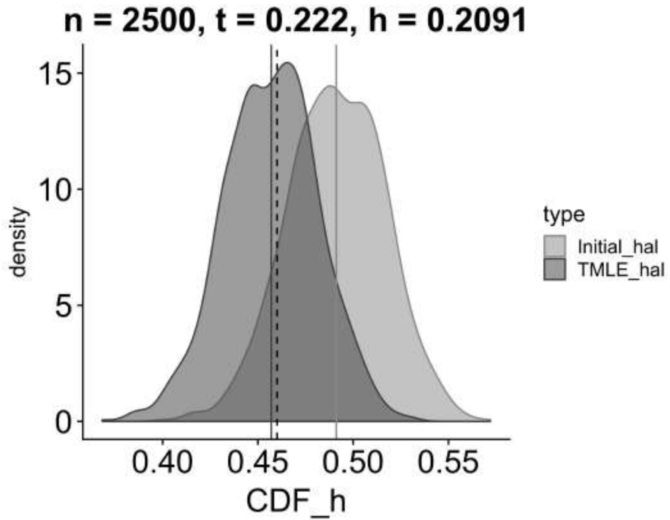
STEP 4:
CONSTRUCT
ESTIMATOR

**STEP 5:
OBTAIN
INFERENCE**

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



Effect of targeting on sampling distribution



Truth = smoothed Pr(TE(W) <= t) at dashed line

What is the optimal rule for assigning corticosteroids to patients in septic shock?

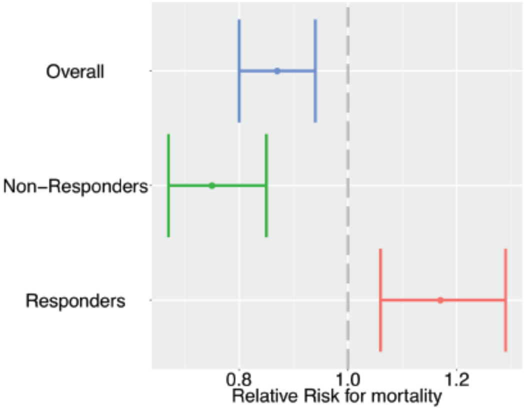
STEP 1:
DESCRIBE
EXPERIMENT

STEP 2:
SPECIFY
STATISTICAL MODEL

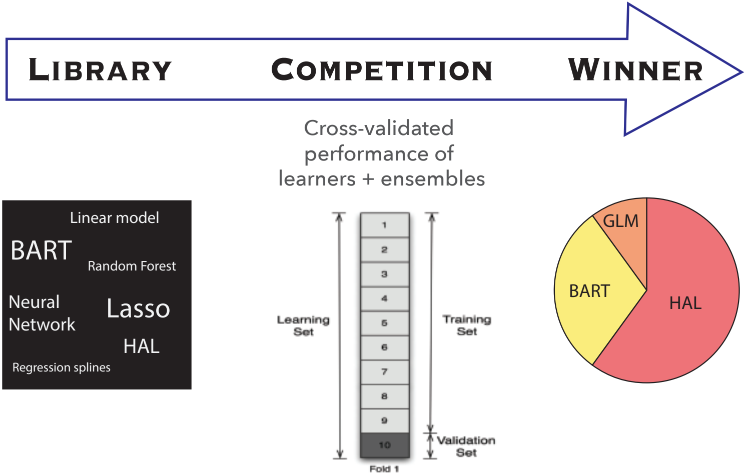
STEP 3:
DEFINE
STATISTICAL QUERY

STEP 4:
CONSTRUCT
ESTIMATOR

STEP 5:
OBTAIN
INFERENCE

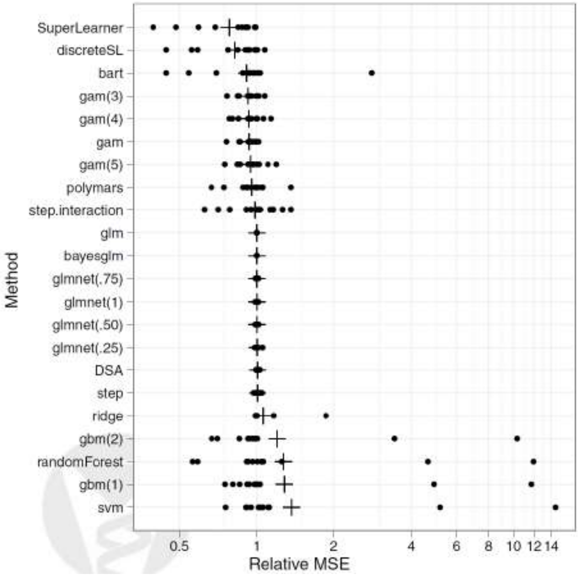


Super learner



Oracle inequality tells us cross-validation is optimal for selection among estimators

Super learner performance in practice



Highly Adaptive Lasso (HAL)

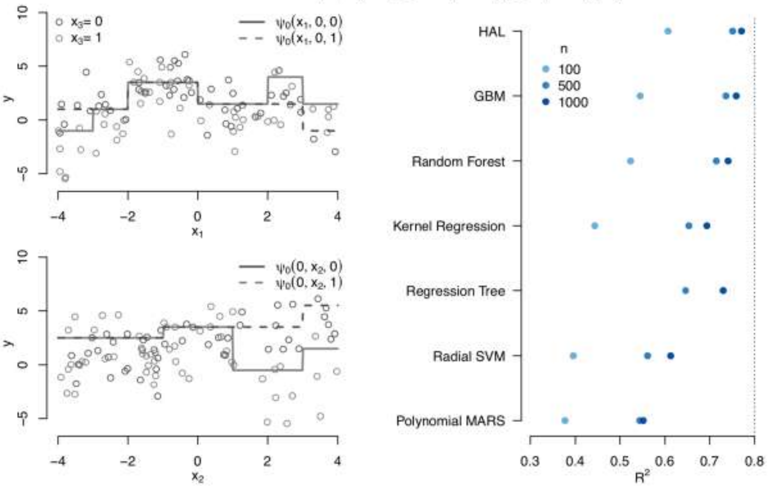
Key Idea

- Any d-dimensional cadlag function (i.e. right-continuous) can be represented as a possibly infinite linear combination of spline basis functions.
- The variation norm / complexity of a function is the L_1-norm of the vector of coefficients.

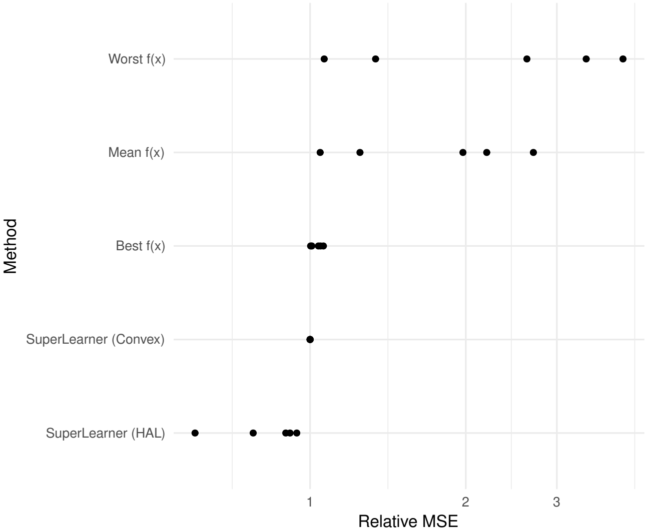
Converges to true function at rate $n^{-1/3}(\log n)^{d/2}$

HAL performance for d=3

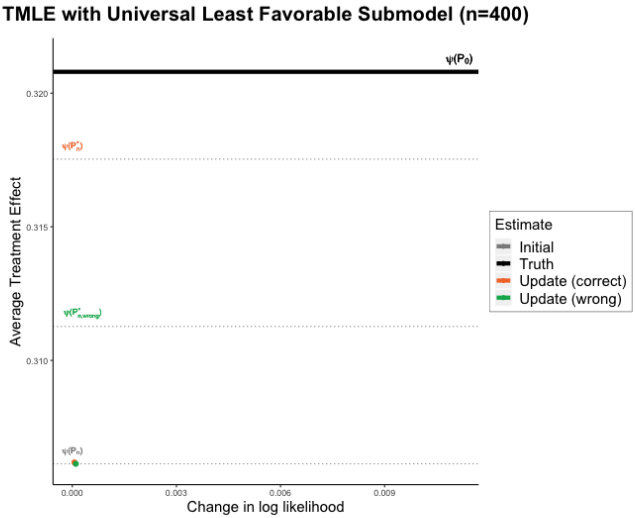
$$\psi_0(x) = -2x_3\, I(x_1 < -3) + 2.5\, I(x_1 > -2) - 2\, I(x_1 > 0) + 2.5x_3\, I(x_1 > 2) - 2.5\, I(x_1 > 3) + I(x_2 > -1) - 4x_3\, I(x_2 > 1) + 2\, I(x_2 > 3)$$



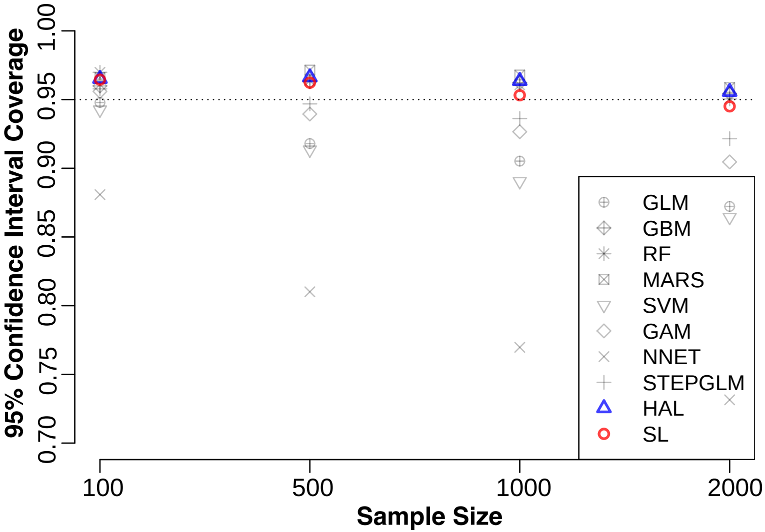
HAL metalearner



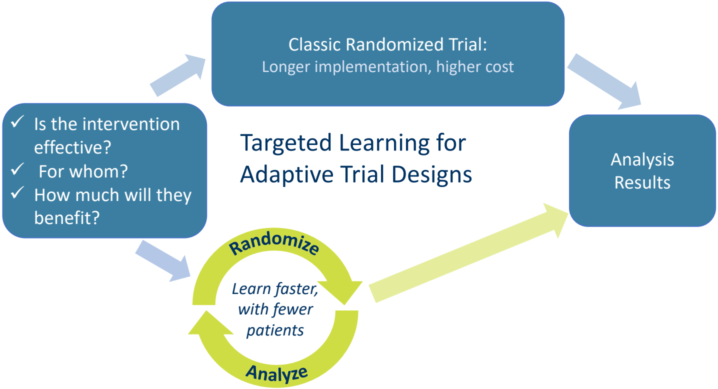
TMLE follows a path of maximal change in target estimand per unit of information/likelihood



Can we break HAL-TMLE?

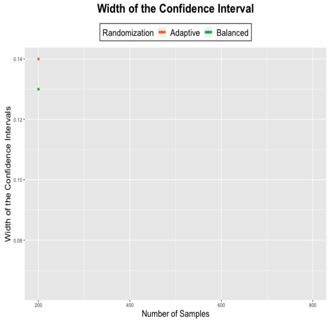


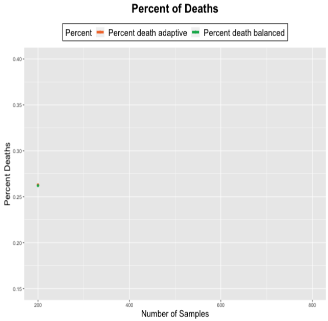
Optimal intervention allocation: “Learn as you go”



Balanced vs. adaptive sequential design

Balanced vs. adaptive sequential design



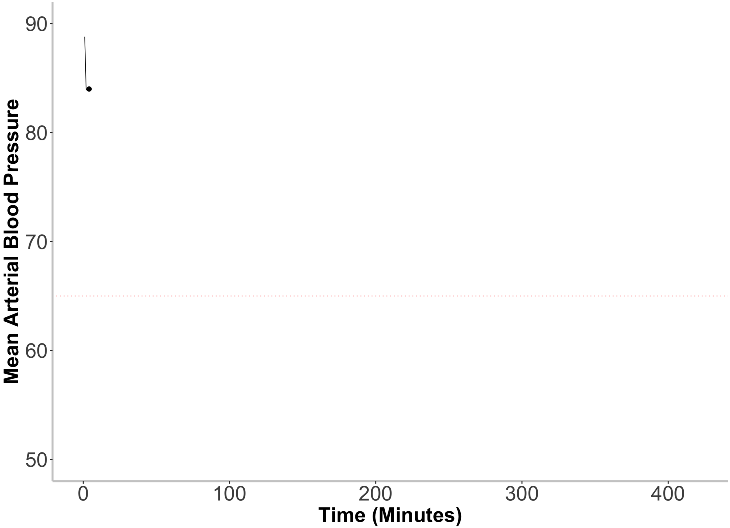


Online super learning in the ICU

Adaptive algorithm

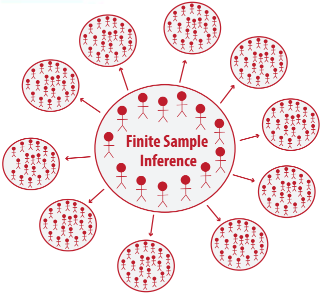
- Regularly updated with batches of new data
- Learns from both
  1. within individual time series, and
  2. across patients
- Uncertainty of forecasts assessed with prediction intervals

15-minute ahead forecasts with prediction intervals for patient with hypotensive episodes



Future of Targeted Learning

U.S. FOOD & DRUG ADMINISTRATION



novo nordisk

accenture