Targeted Machine Learning for Causal Inference Based on Real-World Data

Mark van der Laan
Division of Biostatistics, UC Berkeley

February 5, 2020, FDA Webinar I

Outline

Roadmap of Statistical Learning

- **Observed data**: Realization of a random variable $O^n = (O_1, \ldots, O_n)$ with a probability distribution (say) P_0^n on n units.
- **Model stochastic system of observed data realistically**: Statistical model $\mathcal{M}^n$ is set of possible probability distributions of the data.
- **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. **Estimand is chosen so that it best approximates the answer to causal question of interest**.
- **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.
- **Confidence interval for true answer to query:** Establish approximate sampling probability distribution of the estimator (e.g., based on CLT), and corresponding statistical inference.

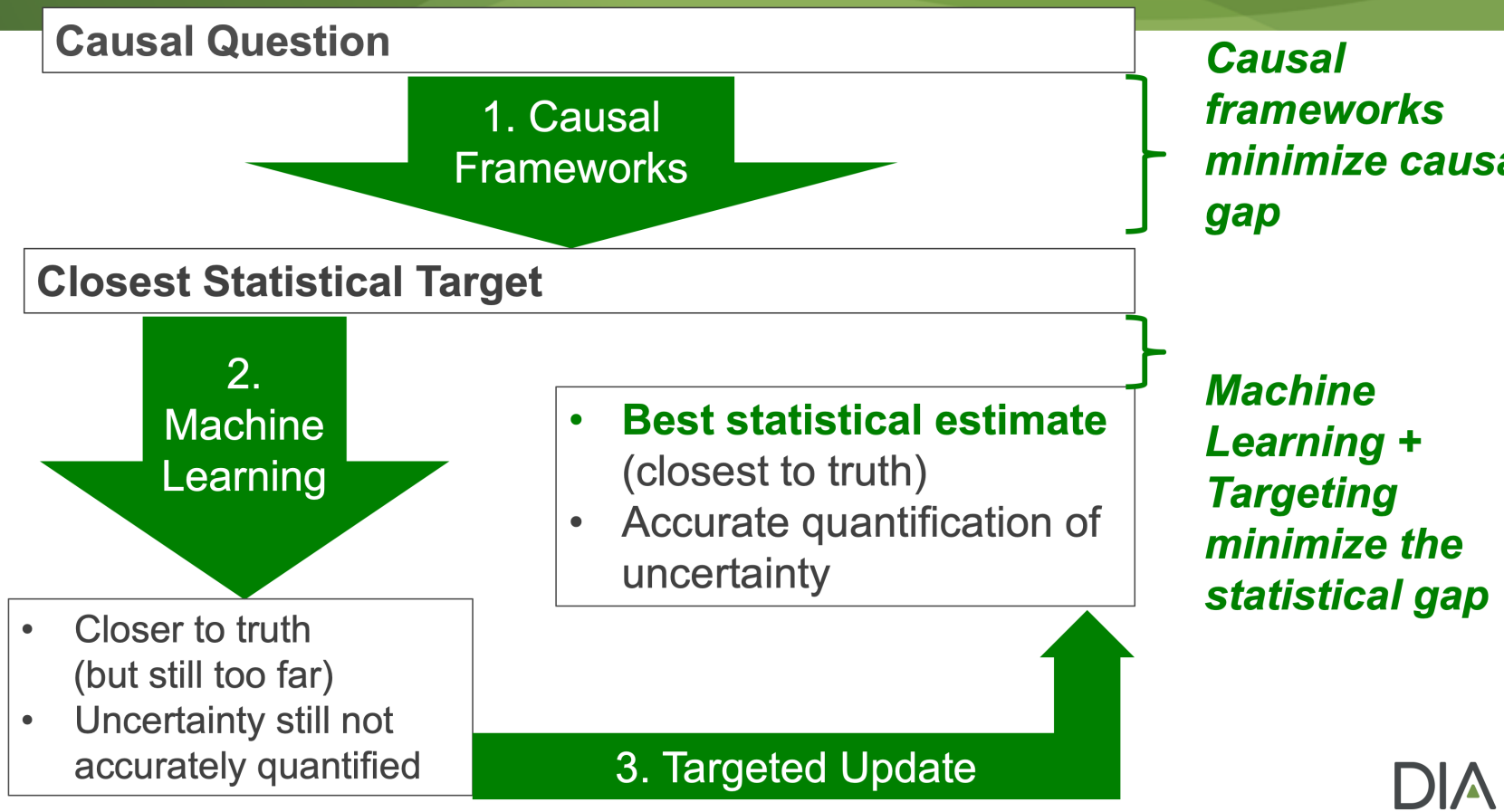
Targeted Learning (TL)

is the subfield of statistics concerned with development of (targeted ML) estimators of the data distribution based on observed data with corresponding plug-in estimates and **confidence intervals** for the desired estimand, **based on realistic statistical models**.

By necessity, TL involves highly data adaptive estimation (i.e., ensemble machine learning) to estimate the relevant complex stochastic relations between the different observed variables.

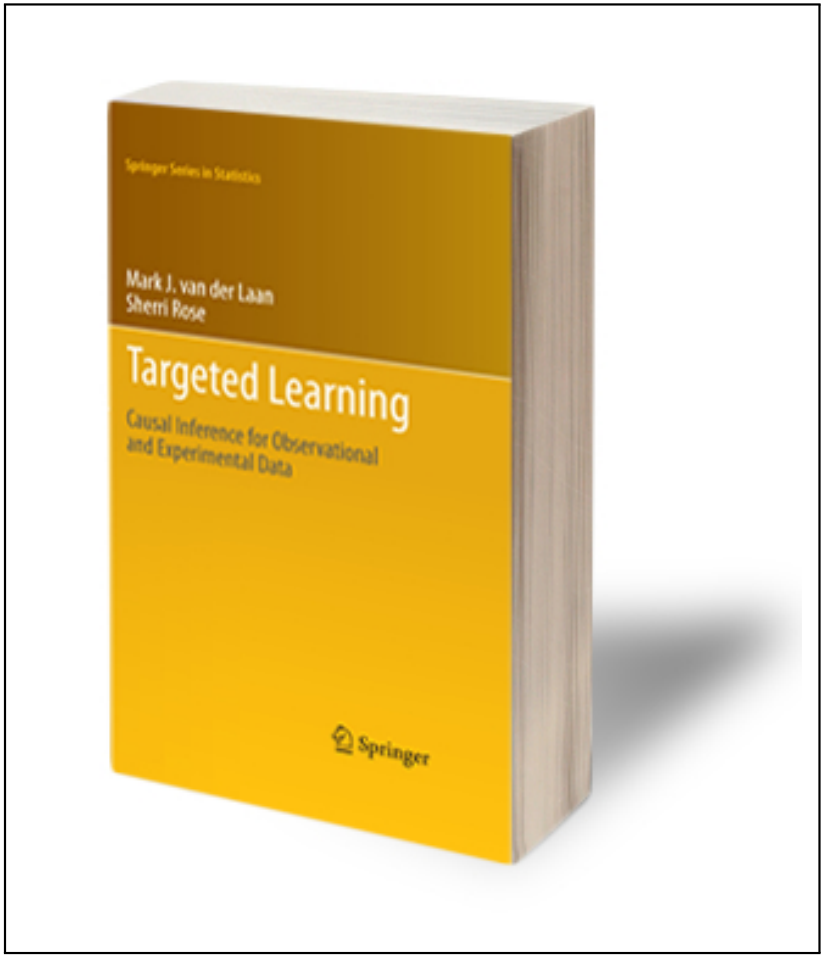
Outline

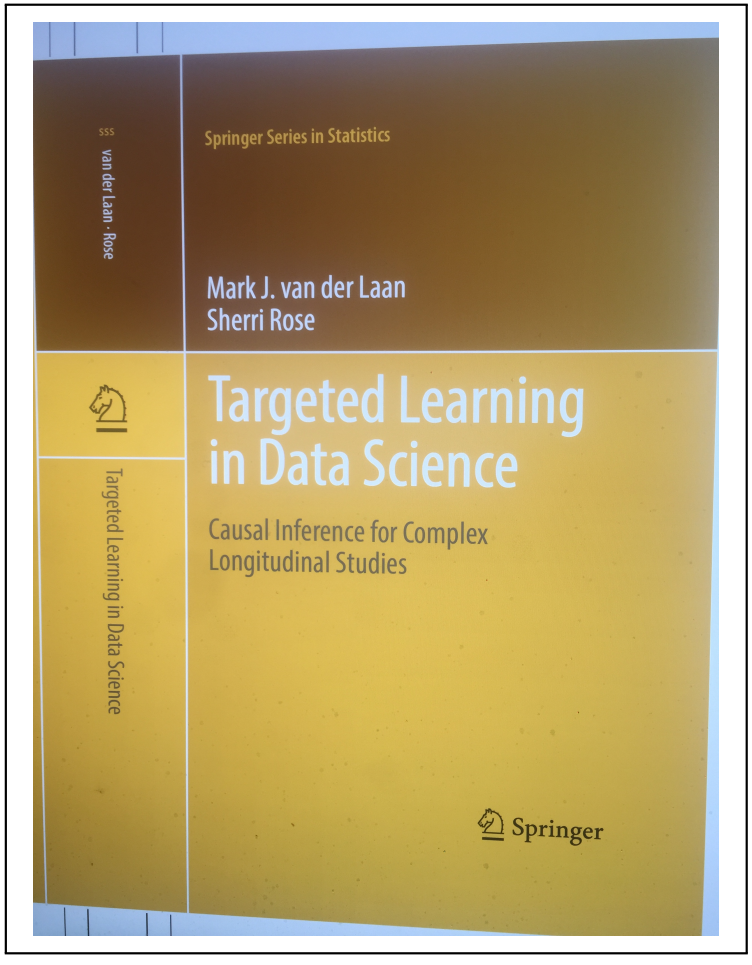
Targeted Learning



DIA

Targeted Learning (https://vanderlaan-lab.org))





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

Outline

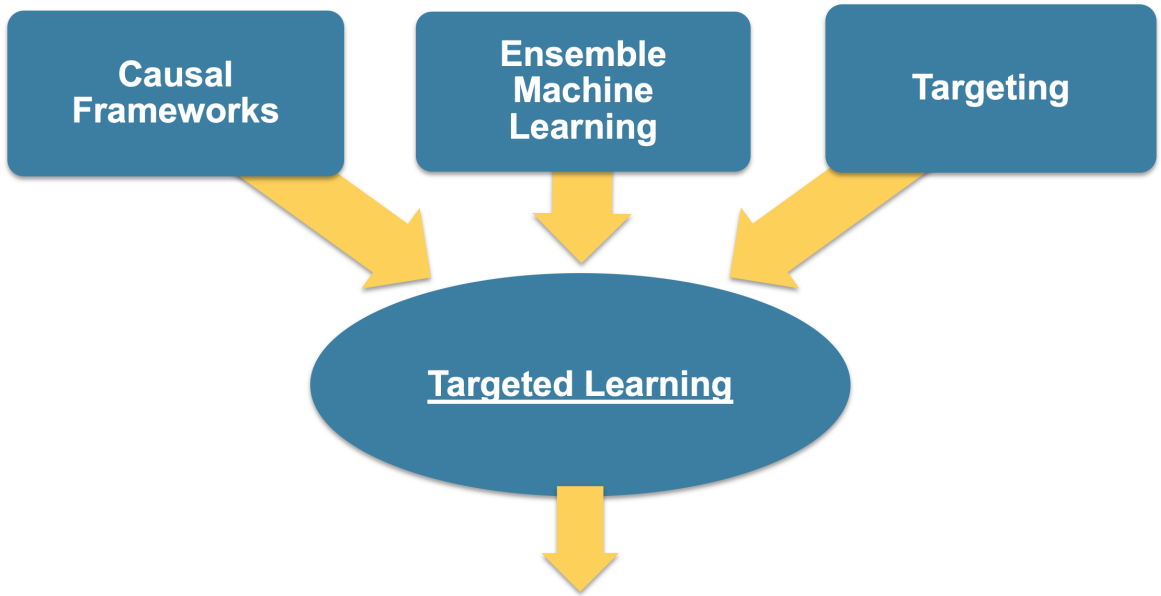
Example: Nonparametric Estimation of Average Treatment Effect

- Unit (i.i.d.) data $O \sim P_0$ consists of baseline covariates W, binary treatment A, and final binary outcome Y.
- Statistical model for the data distribution P_0 is nonparametric.
- Statistical target parameter:

$$\Psi(P) = E_P\{P(Y = 1 \mid A = 1, W) - P(Y = 1 \mid A = 0, W)\}.$$

- Under causal model, randomization assumption, and positivity assumption, $\Psi(P) = E(Y_1 - Y_0)$ is the ATE.
- A TMLE will estimate $P(Y = 1 \mid A, W)$ with (say) **ensemble machine learning** and a subsequent **Targeting step** using logistic regression with off-set initial fit, and clever covariate $(2A - 1)/\hat{P}(A|W)$.

Targeted Learning

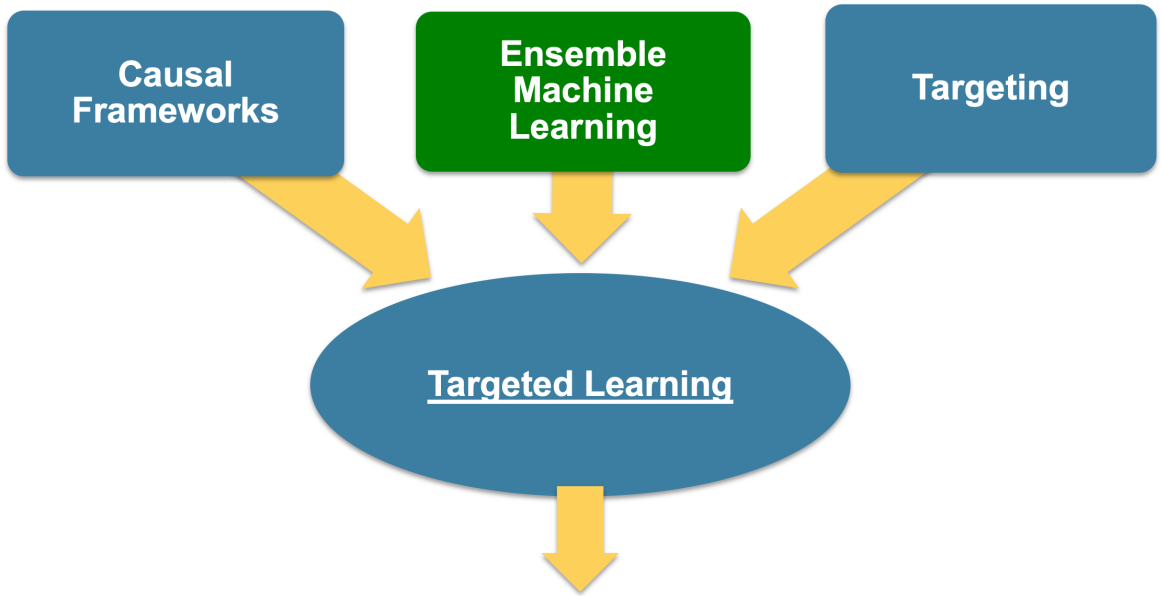


Better (more precise) **answers** to **causal** (actionable) **questions** with **accurate quantification of uncertainty** (signal from noise)

DIA

Outline

Targeted Learning

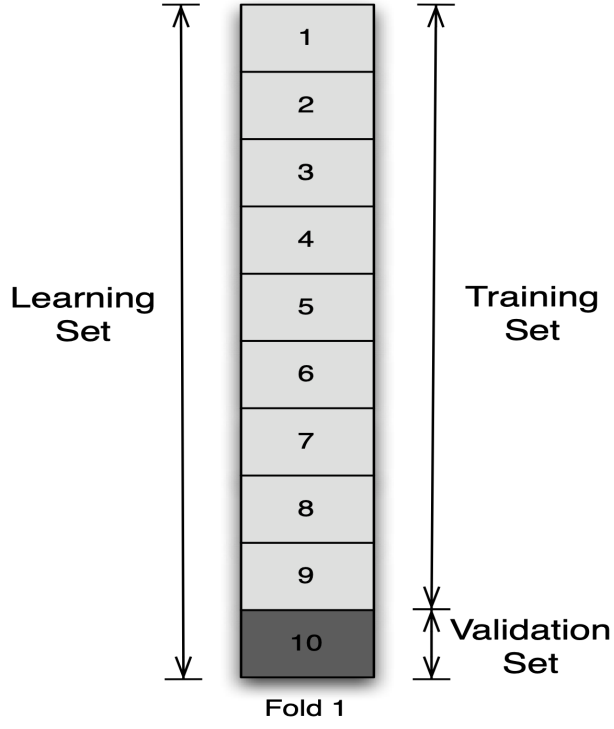


Better (more precise) **answers** to **causal** (actionable) **questions** with **accurate quantification of uncertainty** (signal from noise)

DIA

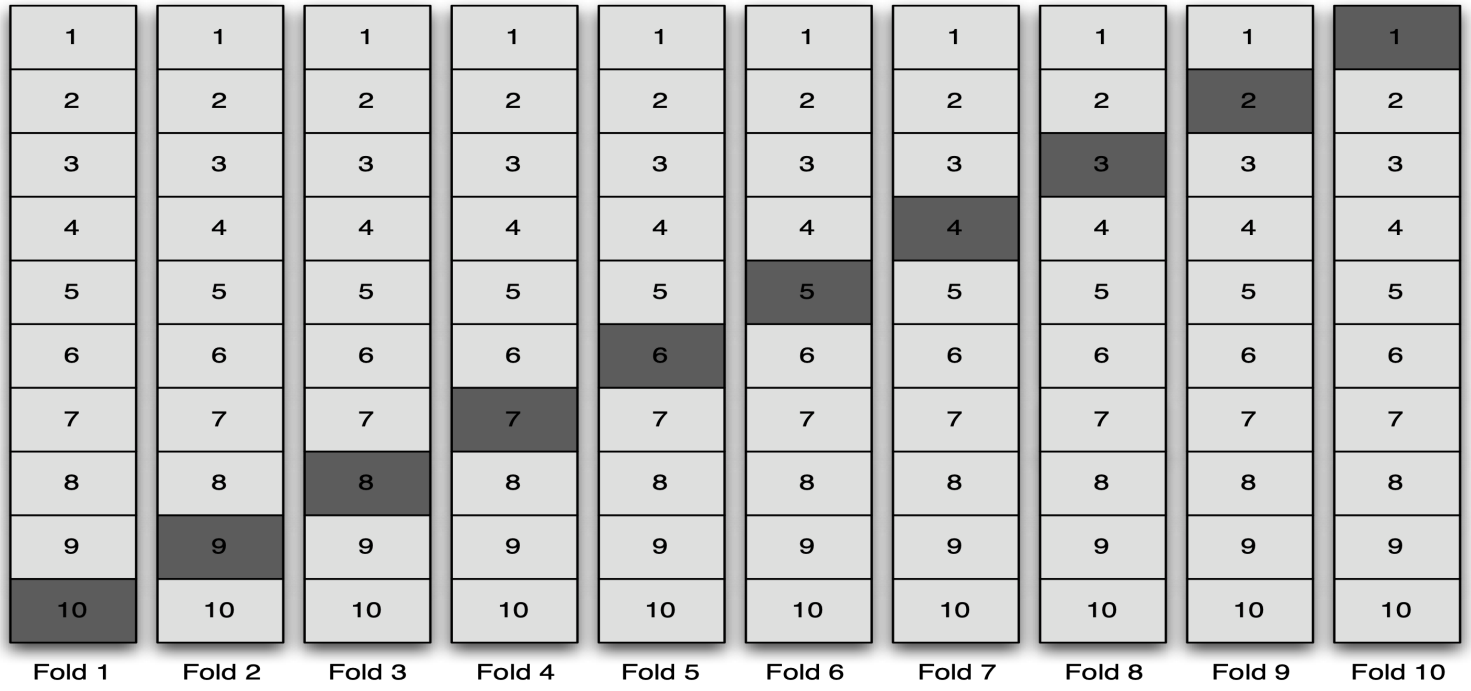
Super Learning: Ensemble Machine Learning

- "Competition" of algorithms
  - Parametric models
  - Data-adaptive (ex. Random forest, Neural nets)
- Best "team" wins
  - Convex combination of algorithms
- Performance judged on independent data
  - V-fold cross validation (Internal data splits)
- Customizable optimality criterion
  - Standard loss function
  - Minimize false negatives with bounded false positives
  - Respect resource constraints



Van der Laan, Polley, 2007

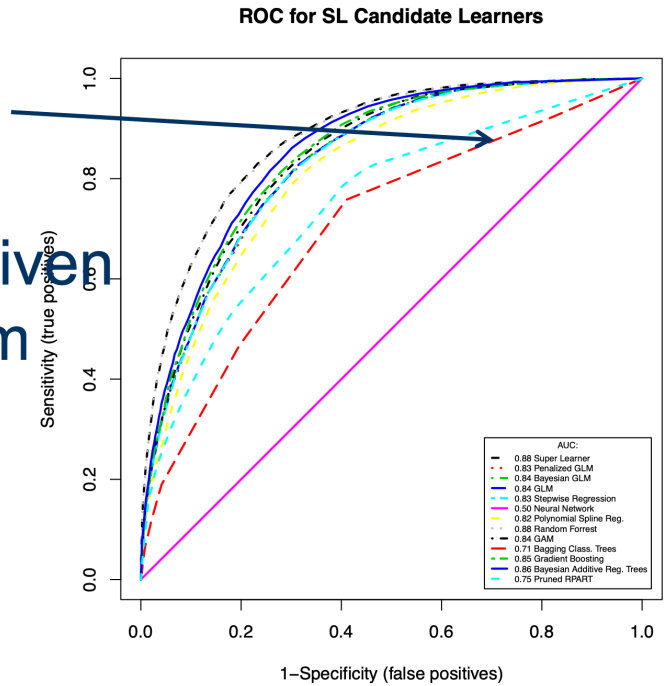
V-fold Cross Validation



DIA

Ex. Mortality prediction for ICU patients

Super Learner:
Best weighted combination of algorithms for a given prediction problem



Pirracchio, Lancet Resp Med 2015

DIA

Cross-validation is optimal for selection among estimators

- We established an oracle inequality for the cross-validation selector among a collection of candidate estimators (e.g, van der Laan, Dudoit, 03, van der Vaart et al 06).
- Oracle selector chooses the estimator closest to the true function w.r.t. loss-based dissimilarity.
- It establishes that the loss-based dissimilarity with truth of the cross-validated selected estimator divided by the loss-based dissimilarity of the oracle selected estimator converges to 1, even as the number of candidate estimators converges to infinity as a polynomial in sample size.
- Only condition is that loss-function is uniformly bounded.

Highly Adaptive Lasso (HAL)

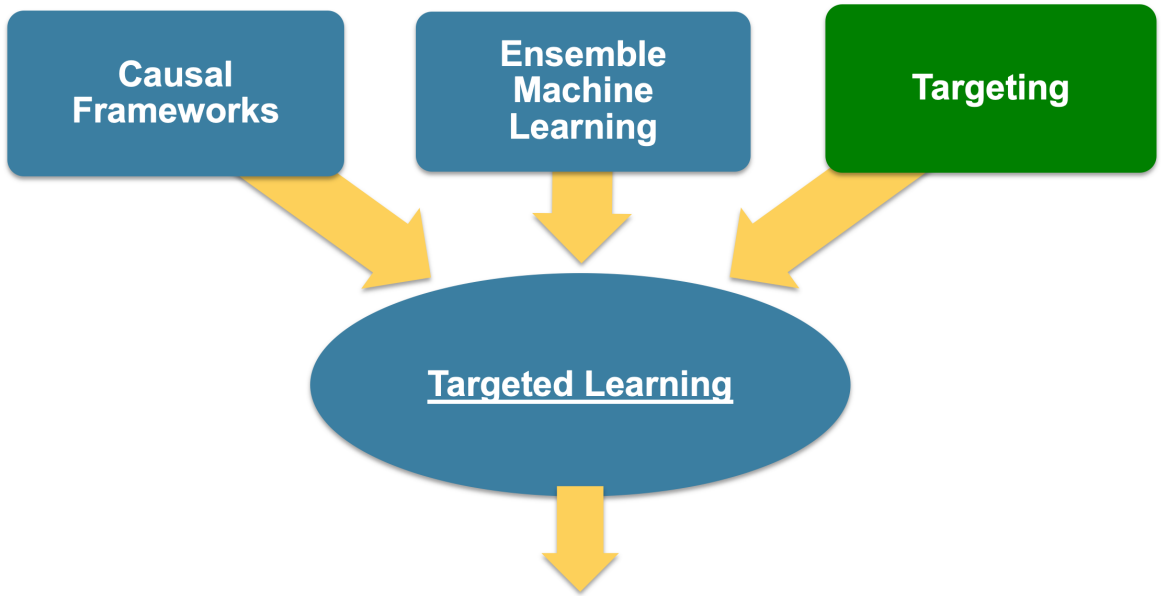
- This is a machine learning algorithm that estimates functionals (e.g outcome regression and propensity score) by approximating them with linear model in many ($\leq n2^d$) tensor product indicator basis functions, constraining the L_1-norm of the coefficient vector, and choosing it with cross-validation (vdL, 2015, Benkeser, vdL, 2017)
- Guaranteed to converge to truth at rate $n^{-1/3}(\log n)^d$ in sample size n (Bibaut, vdL, 2019): only assumption that true function is right-continuous, left-hand limits, and has finite sectional variation norm.
- When used in super-learner library (or by itself), TMLE (targeted learning) is guaranteed **consistent, (double robust) asymptotically normal and efficient**: one only needs to assume *strong positivity assumption*.

Example: HAL-MLE of conditional hazard

- Suppose that $O = (W, A, \tilde{T} = \min(T, C), \Delta = I(T \leq C))$, and that we are interested in estimating the conditional hazard $\lambda(t \mid A, W)$.
- Let $L(\lambda)$ be the log-likelihood loss.
- If T is continuous, we could parametrize $\lambda(t \mid A, W) = \exp(\psi(t, A, W))$, or, if T is discrete, $\text{Logit}\lambda(t \mid A, W) = \psi(t, A, W)$.
- We can represent $\psi = \sum_{s \subset \{1,\ldots,d\}} \beta_{s,j} \phi_{u_{s,j}}$ as linear combination of indicator basis functions $\phi_{u_{s,j}}(x) = I(x_s \geq u_{s,j})$, where L^1-norm of β represents the sectional variation norm of ψ.
- Therefore, we can compute the HAL-MLE of λ with either Cox-Lasso or logistic Lasso regression (e.g., glmnet() using family is Cox).

Outline

Targeted Learning



Better (more precise) **answers** to **causal** (actionable) **questions** with **accurate quantification of uncertainty** (signal from noise)

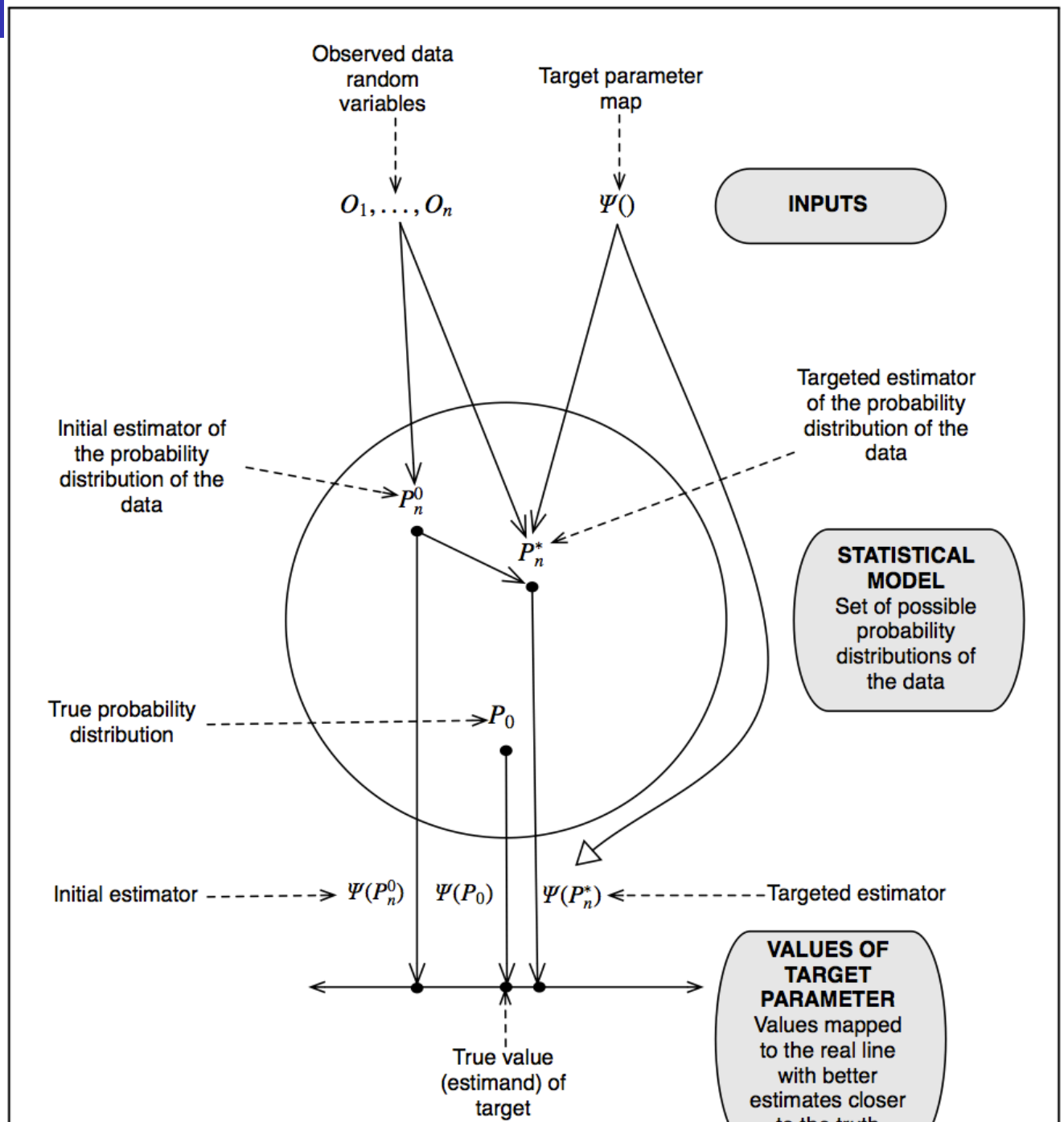
DIA

Targeted Update of Machine Learning

- Don't try to do a good job for all questions at once.
- Focus estimation where it matters most for question at hand.

1. Less bias (closer to truth).
2. Sampling distribution approximately normal, more accurate quantification of uncertainty.

Targeted minimum loss based estimation (TMLE)



Targeted Minimum Loss Based Estimation (TMLE)

- Super learning provides an initial estimator $\mathbf{P}_n$ of stochastic system P_0.
- Determine mathematically the fluctuation strategy (least favorable submodel) $\mathbf{P}_{n,\epsilon}$ of the super-learner fit $\mathbf{P}_n$ with tuning parameter ϵ **so that a small change in ϵ corresponds with a maximal small change** in estimated answer $\Psi(\mathbf{P}_{n,\epsilon})$ to query $\Psi(P_0)$: i.e., score equals canonical gradient/**efficient influence curve** $D^*(\mathbf{P}_n)$.
- Determine the optimal amount ϵ_n of fluctuation based on the data (e.g., maximum likelihood estimation).
- The resulting update $\mathbf{P}_n^* = \mathbf{P}_{n,\epsilon_n}$ of the initial estimator of stochastic system is the TMLE of P_0 and it implies the TMLE $\Psi(\mathbf{P}_n^*)$ of the answer to query.
- Thanks to TMLE-update, TMLE solves optimal score equation $P_n D^*(\mathbf{P}_n^*) \approx 0$, and is asymptotically normally distributed around true answer to query with minimal asymptotic variance.

Three general methods for efficient estimation in literature

Three general methods result in asymptotically efficient estimators, given good initial estimator $\mathbf{P}_n$ of data distribution P_0, using canonical gradient $D^*(P)$ of target estimand as ingredient:

- **One-step estimator:** $\psi_n^1 = \Psi(\mathbf{P}_n) + P_n D^*(\mathbf{P}_n)$.
- **Estimating equation estimator:** Assume estimating function representation $D^*(P) = D^*(\psi, \eta(P))$; let ψ_n solution of $P_n D^*(\psi, \eta(\mathbf{P}_n)) = 0$.
- **TMLE:** $\mathbf{P}_{n,\epsilon}$ least favorable submodel through initial $\mathbf{P}_n$; ϵ_n MLE; $P_n^* = \mathbf{P}_{n,\epsilon_n}$; TMLE is $\Psi(P_n^*)$.
- TMLE is general method that updates initial $\mathbf{P}_n$ into improved fit $\mathbf{P}_n^*$ that solves **user supplied set of equations** $P_n D(\mathbf{P}_n^*) \approx 0$, allowing for various additional statistical properties beyond asymptotic efficiency.

Each one of the methods has a sample splitting analogue removing Donsker class condition.

Outline

Simulation design for ATE example

We repeatedly sampled random data generating mechanisms and simulated samples of size $n \in \{100, 500, 1000, 2000\}$ for a total of 25,000 different data generating mechanisms of (W, A, Y).

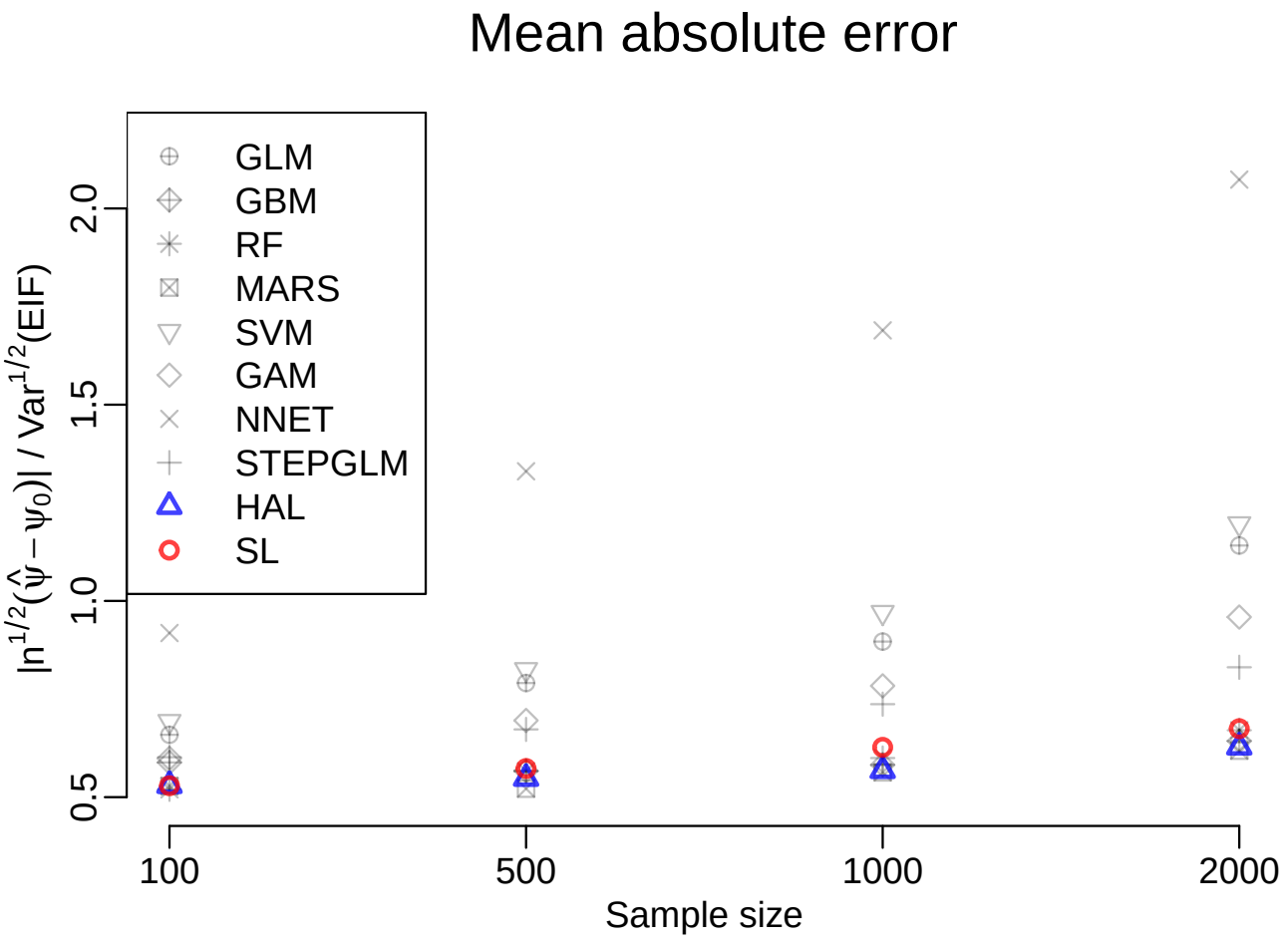
We computed TMLEs of the ATE based on different estimators of $E_0(Y \mid A, W)$ and $P_0(A = 1 \mid W)$.

- GLM, Bayes GLM, stepwise GLM (AIC), stepwise GLM (p-value), stepwise GLM with two-way interactions, intercept-only GLM, GAM, GBM*, random forest*, linear SVM*, neural nets*, regression trees*, HAL
- Super Learner (based on these algorithms)
- * = tuning parameters selected via cross-validation

Estimators compared on their absolute error (relative to best achievable SE) and coverage probability of 95% oracle confidence intervals.

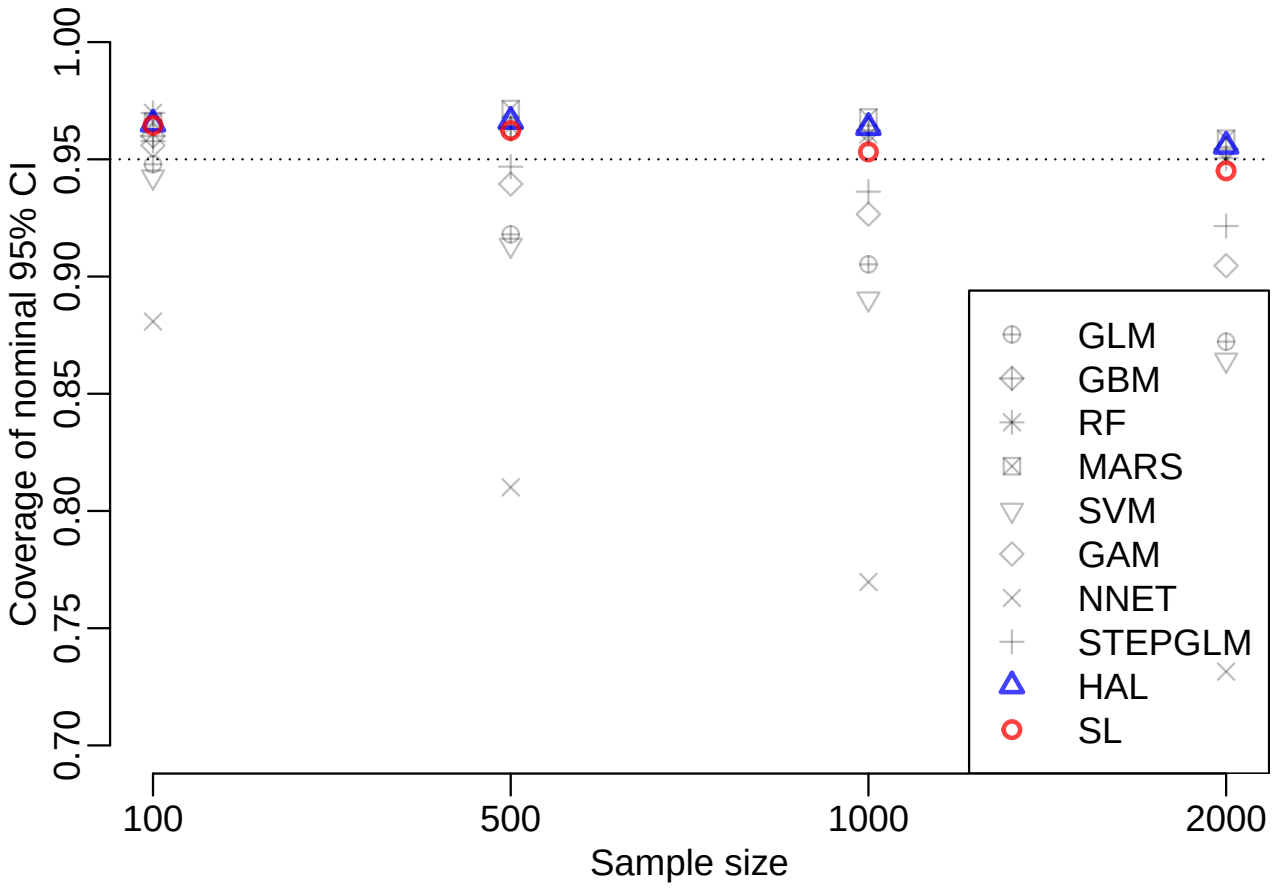
Results – absolute error by sample size

HAL-TMLE exhibited excellent accuracy relative to competitors.



Results – coverage by sample size

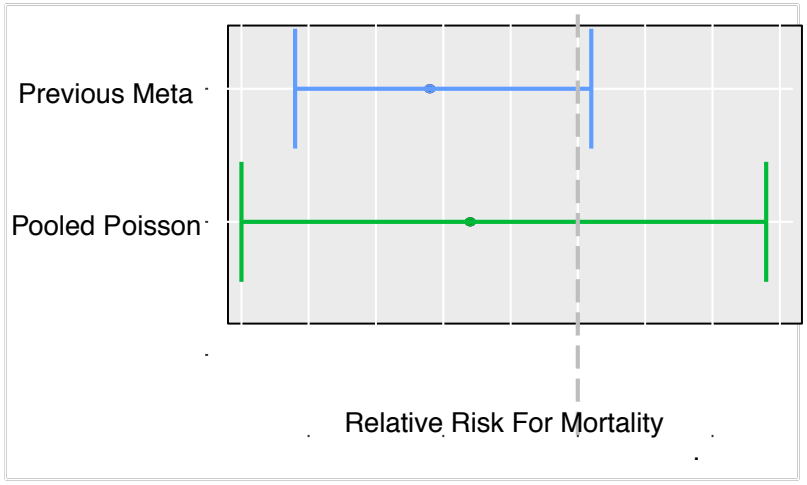
HAL-TMLE achieves approximate Normality in reasonable sample sizes.



Outline

1 Better, cheaper trials

Do corticosteroids reduce mortality for adults with septic shock?



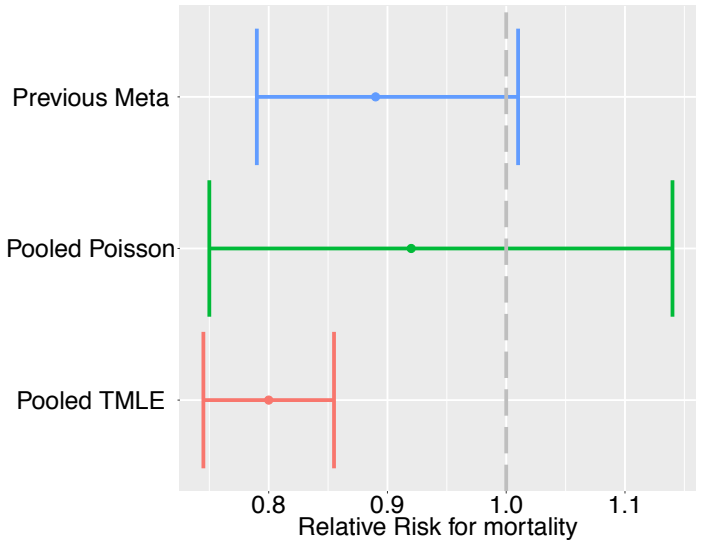
Previous Meta-Analysis of 31 trials: ***No significant benefit***

Pooled analysis of 3 major RCTs (1300 patients) with standard methods: ***No significant benefit***

Pirracchio 2016

Better, cheaper trials

Do corticosteroids reduce mortality for adults with septic shock?



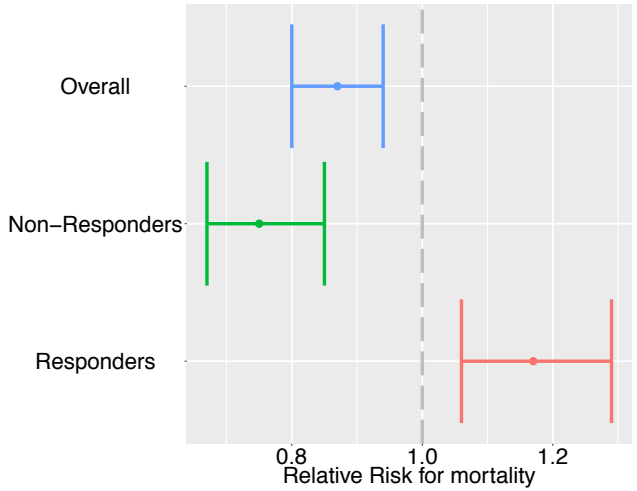
Previous Meta-Analysis of 31 trials: No significant benefit

Pooled analysis of 3 major RCTs (1300 patients) with standard methods: **No significant benefit**

Pooled analysis of 3 major RCTs using Targeted Learning: significant reduction of mortality.

Not just is there an effect, but for whom?

- In Sepsis re-analysis: Targeted Learning showed **all benefit** occurred in a key subgroup
  - Heterogeneity in patient populations one cause of inconsistent results

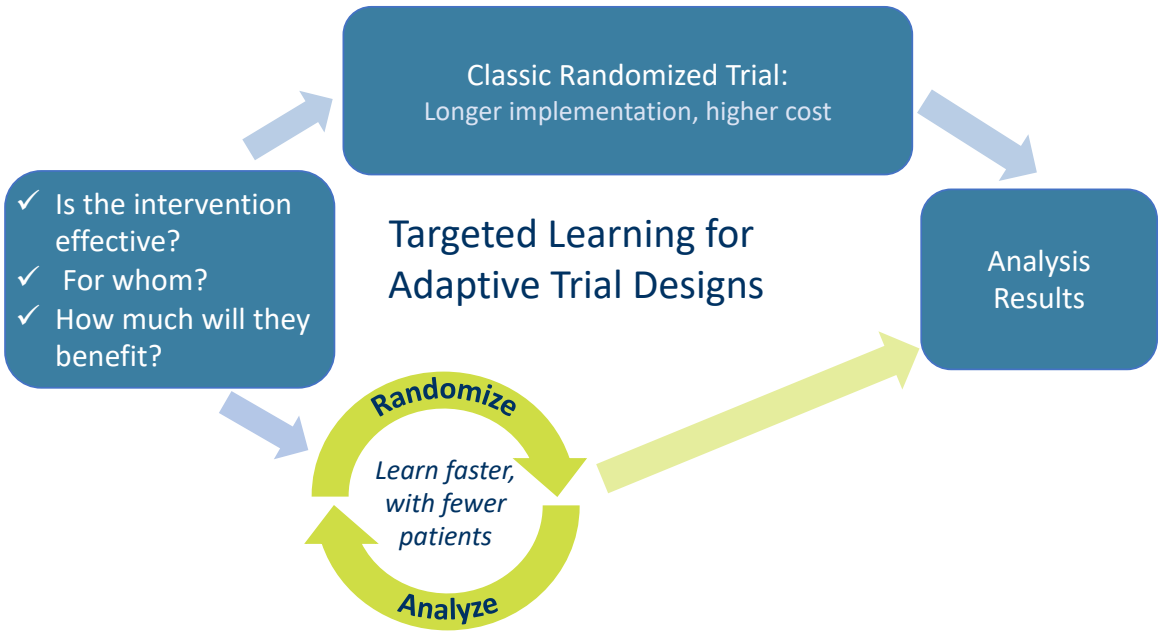


Effect Heterogeneity by Response to ACTH Stimulation

Outline

Robust estimation and inference for sequential designs adapting intervention allocation probabilities based on learning from past

Optimal intervention allocation: "Learn as you go"



Outline

One-step TMLE of treatment specific survival curve

Data structure

- $O = (W, A, \tilde{T} = \min(T, C), \Delta = I(T \leq C))$.
- dynamic treatment intervention: $W \rightarrow d(W)$.
- $S_d(t)$ is defined by

$$\Psi(P)(t) = E_P\left[P\left(T > t | A = d(W), W\right)\right]$$

- Focus on $d(W) = 1$.

Efficient influence curve

The efficient influence curve for $\Psi(P)(t)$ is (Hubbard et al., 2000)

$$D_t^*(P) = \sum_{k \leqslant t} h_t(g_A, S_{A_c}, S)(k, A, W) \Big[I(\widetilde{T} = k, \Delta = 1) -$$

$$I(\widetilde{T} \geqslant k) \lambda(k | A = 1, W) \Big] + S(t | A = 1, W) - \Psi(P)(t)$$

$$\equiv D_{1,t}^*(g_A, S_{A_c}, S) + D_{2,t}^*(P),$$

where

$$h_t(g_A, S_{A_c}, S)(k, A, W) =$$

$$- \frac{I(A = 1) I(k \leqslant t)}{g_A(A = 1 | W) S_{A_c}(k\_ | A, W)} \frac{S(t | A, W)}{S(k | A, W)}.$$

From local least favorable submodel to universal least favorable submodel

- A local least favorable submodel (LLFM) for $S_d(t)$ around initial estimator of conditional hazard:

$$\text{logit}(\lambda_{n,\varepsilon}(\cdot|A=1,W)) = \text{logit}(\lambda_n(\cdot|A=1,W)) + \varepsilon h_t.$$

- Similarly, we have this local least favorable submodel for a vector $(S_d(t) : t)$ by adding vector $(h_t : t)$ extension.
- These imply universal least favorable submodels for single survival probability and for whole survival function.

Outline

General Longitudinal Data Structure for Complex Observational Studies

We observe n i.i.d. copies of a longitudinal data structure

$$O = (L(0), A(0), \ldots, L(K), A(K), Y = L(K+1)),$$

where $A(t)$ denotes a discrete valued **intervention node** whose effect we desire to evaluate, $L(t)$ is an intermediate covariate and outcome realized in between intervention nodes $A(t-1)$ and $A(t)$, $t = 0, \ldots, K$, and Y is a final **outcome** of interest.

Survival outcome example: For example,

$$\begin{aligned}
A(t) &= (A_1(t), A_2(t)) \\
A_1(t) &= \text{Indicator of being treated at time } t \\
A_2(t) &= \text{Indicator of being right-censored at time } t \\
Y(t) &= \text{Indicator of observing a failure by time } t \\
L(t) & \quad \text{Vector of time-dependent measurements} \\
Y(t) &\subset L(t) \text{and } Y = Y(K+1).
\end{aligned}$$

Outline

A real-world CER study comparing different rules for treatment intensification for diabetes

- Data extracted from diabetes registries of 7 HMO research network sites:
  - Kaiser Permanente
  - Group Health Cooperative
  - HealthPartners
- Enrollment period: Jan 1^{st} 2001 to Jun 30^{th} 2009

Enrollment criteria:

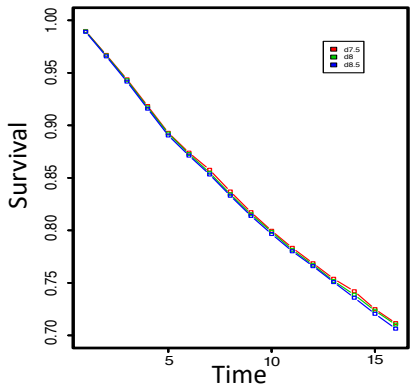
- past A1c$<$ 7% (glucose level) while on 2+ oral agents or basal insulin
- 7% $\leq$ latest A1c $\leq$ 8.5% (study entry when glycemia was no longer reined in)

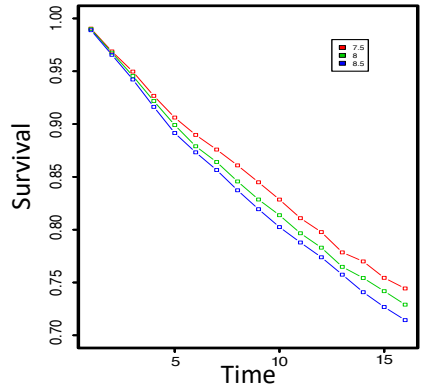
Longitudinal data

- Follow-up til the earliest of Jun 30^{th} 2010, death, health plan disenrollment, or the failure date
- Failure defined as onset/progression of albuminuria (a microvascular complication)
- Treatment is the indicator being on "treatment intensification" (TI)
- $n \approx 51,000$ with a median follow-up of 2.5 years.

2 Better clinical decisions from observational data

- When to intensify treatment for diabetic patients?
  - How best to respond to a series of biomarker measures?
  - 51,000 patients from 7 HMO diabetes registries





Standard methods: No benefit to more aggressive intensification strategy

Targeted Learning: More aggressive intensification protocols result in better outcomes

Neugebauer et al, 2014

Outline

Inference with TMLE

- TMLE is **asymptotically linear with influence curve the canonical gradient**, so that Wald-type confidence intervals are based on estimating variance of its influence curve.
- The simple sample variance of influence curve can underestimate the variance if initial estimator is very adaptive or lack of positivity.
- Robust estimation of this variance by using sample splitting, or TMLE plug-in estimator corrects for this finite sample bias, and can be important (Tran et al, 19).
- One can also use the nonparametric bootstrap if one uses HAL as initial estimator (Cai, vdL, 19), resulting in better finite sample coverage by also picking up higher order behavior.

Outline

Advancing the vanilla TMLE: C-TMLE and extra targeting

- The least favorable parametric fluctuation model often depends on nuisance parameter (e.g., propensity score).
- C-TMLE targets estimation of this nuisance parameter based on criterion how well TMLE fits target estimand.
- Important for observational studies (vdL, Gruber, 2010 etc).
- By adding additional parameters to fluctuation model TMLE solves additional score equations that can be chosen to target second order remainder, and thereby improve finite sample performance.
- This has resulted in higher-order TMLE, double robust inference TMLE, etc (vdL, 14, Benkeser et al., Carone et al).

Outline

Preparing SAP based on TMLE

- **Prior** data or **outcome blind** data can be used to decide on **target estimand** supported by data.
- Prior data can also be used to set up **realistic simulation** to benchmark *specifications* of TMLE implementation, where benchmarks includes confidence interval coverage and type I error control.
- These **specifications of TMLE** include deciding on library of SL; sample splitting version; C-TMLE for nuisance parameter; adaptive truncation; TMLE-update step (e.g, possible extra targeting).
- Once one commits, it freezes the **a priori-specified estimator** that can be submitted as part of SAP for FDA approval.

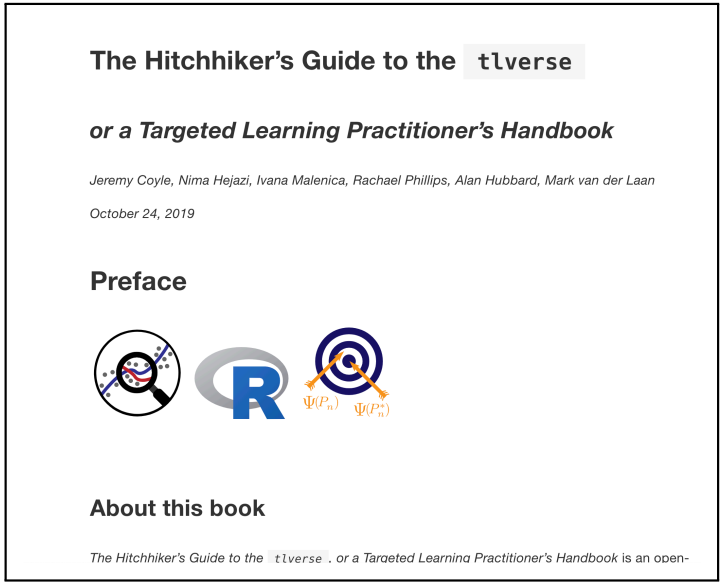
Outline

Concluding Remarks

- **Targeted Learning** *optimally estimates* a causal impact of multiple time point intervention on an outcome for complex observational or randomized studies involving *random monitoring*, *time-dependent treatment*, *informative drop-out and missingness*.
- It integrates **causal inference**, **machine learning**, **statistical theory**.
- The estimate is accompanied with **confidence interval and p-value** for testing.
- The possible *gap between the estimand and causal quantity* due to violation of identification assumptions (e.g., no unmeasured confounder assumption) can be acknowledged through a **sensitivity analysis**.
- Beyond a large collection of R-packages, we also developed a unifiying structured targeted learning software environment **tlverse()** with growing number of tools.

`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