Targeted Learning with Application in Rare Disease Trials

Mark van der Laan

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

8th New England Rare Disease Statistics (NERDS),
October 10, 2025

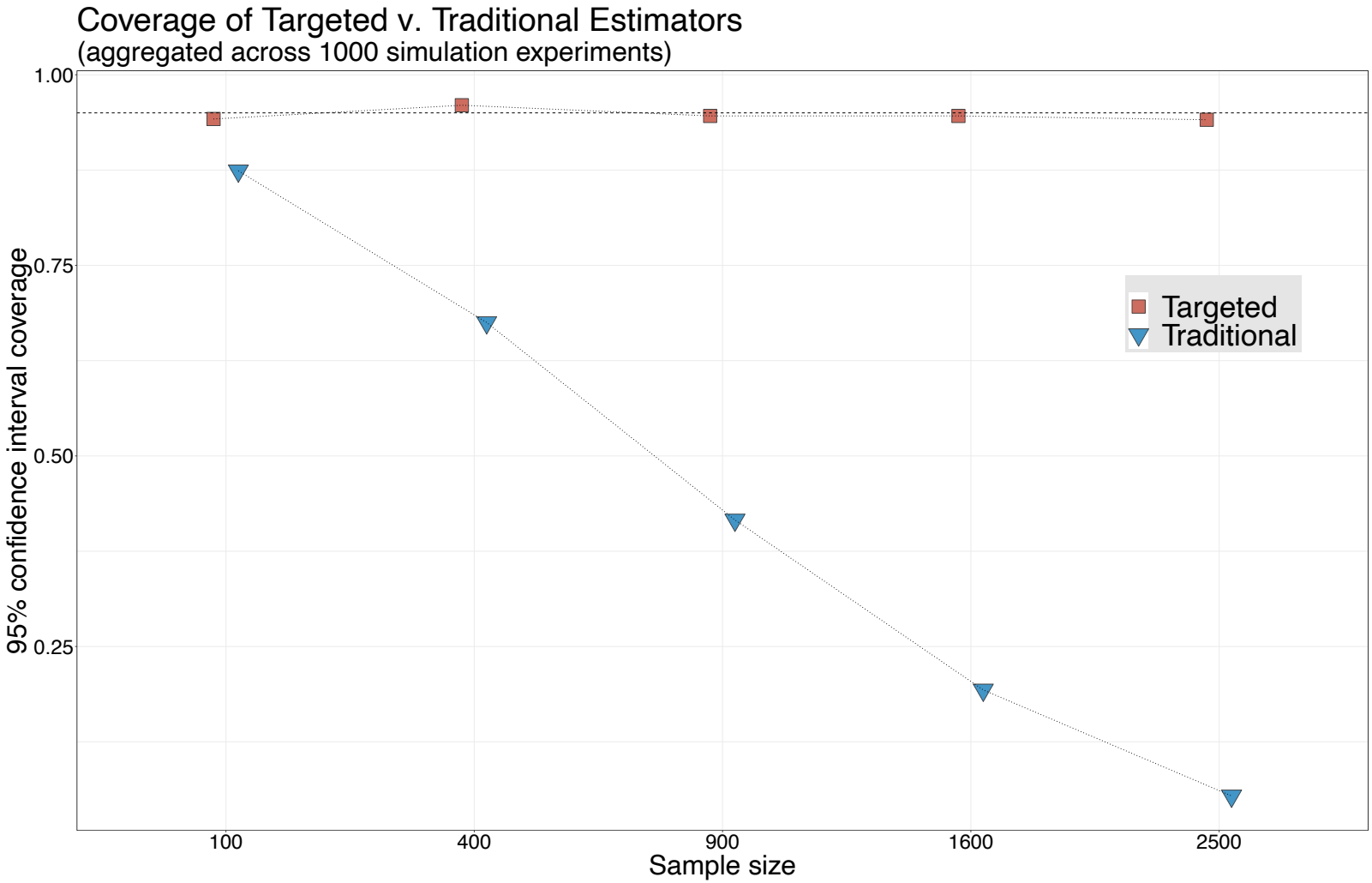
Acknowledgements: Rachael Phillips, Tianyue Zhou, Susan Gruber, Ivana Malenica, Sky Qiu, Lei Nie,

Wonyul Lee, Hana Lee

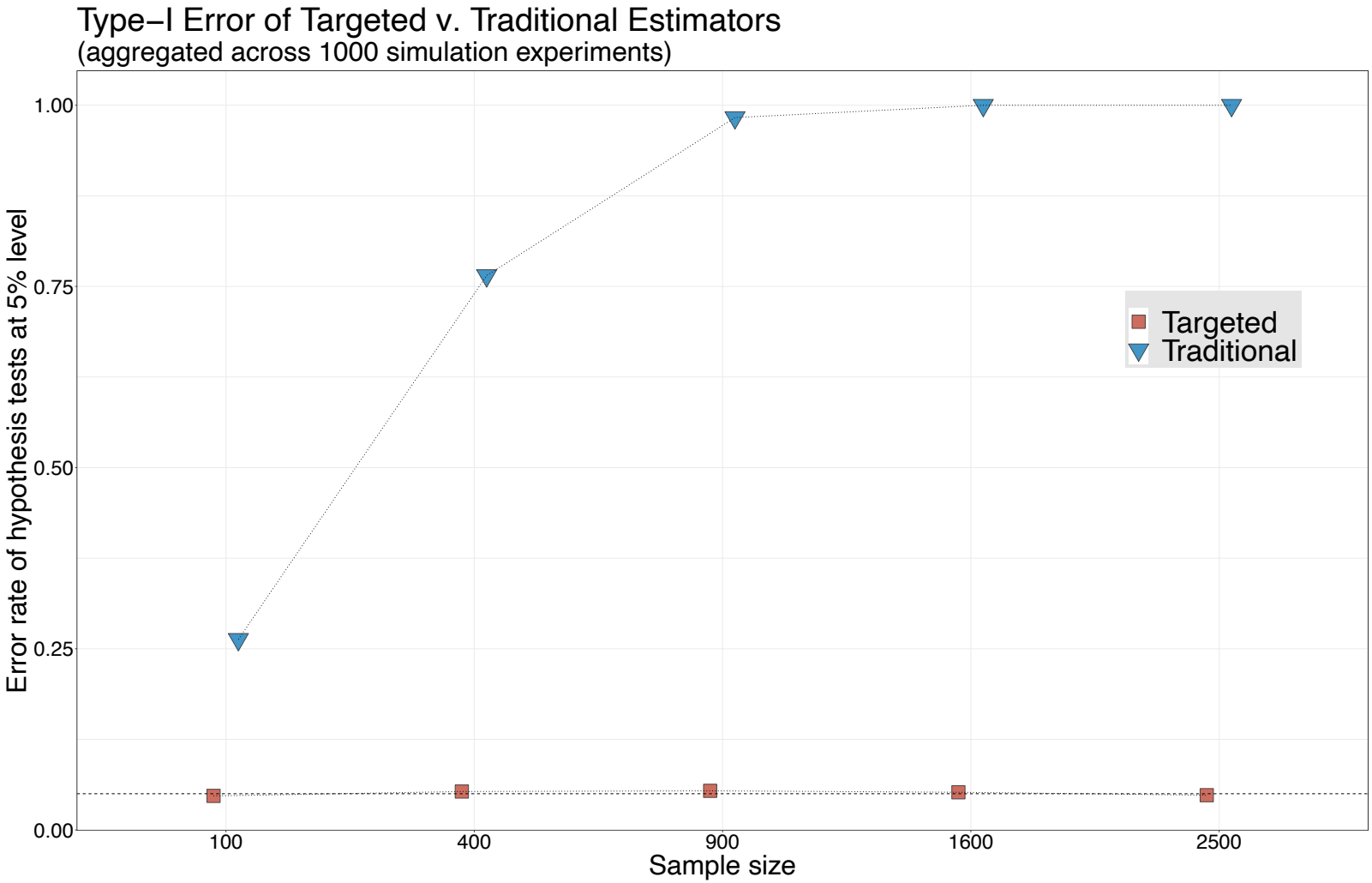
Traditional toolbox for statistics: Recipe oriented, enforces false constraints, not made for Big Data

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

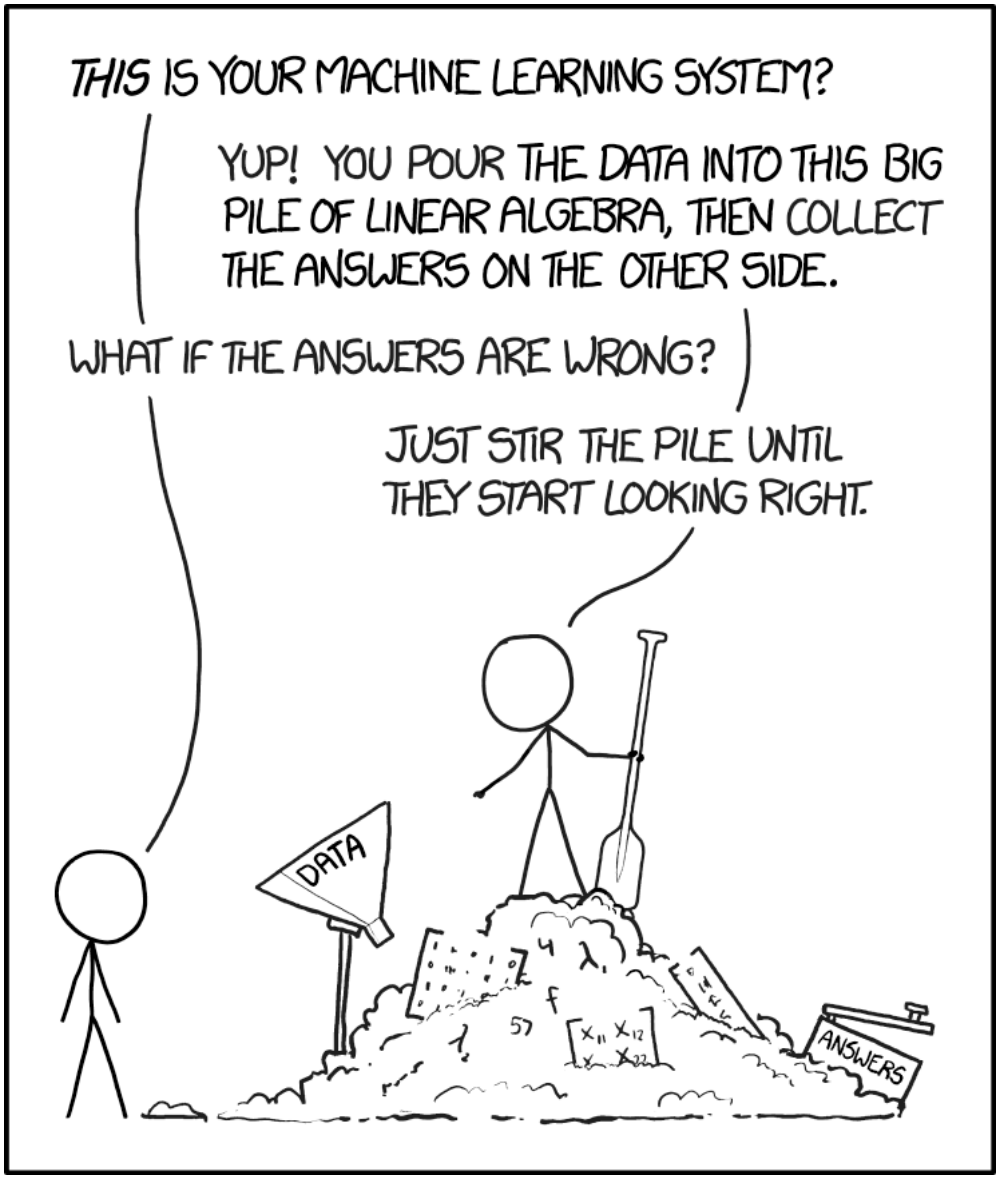
Performance of traditional tools: Coverage of Confidence Intervals deteriorates with sample size



Performance of traditional tools: Type I error deteriorates with sample size



Traditional tools invite/encourage 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

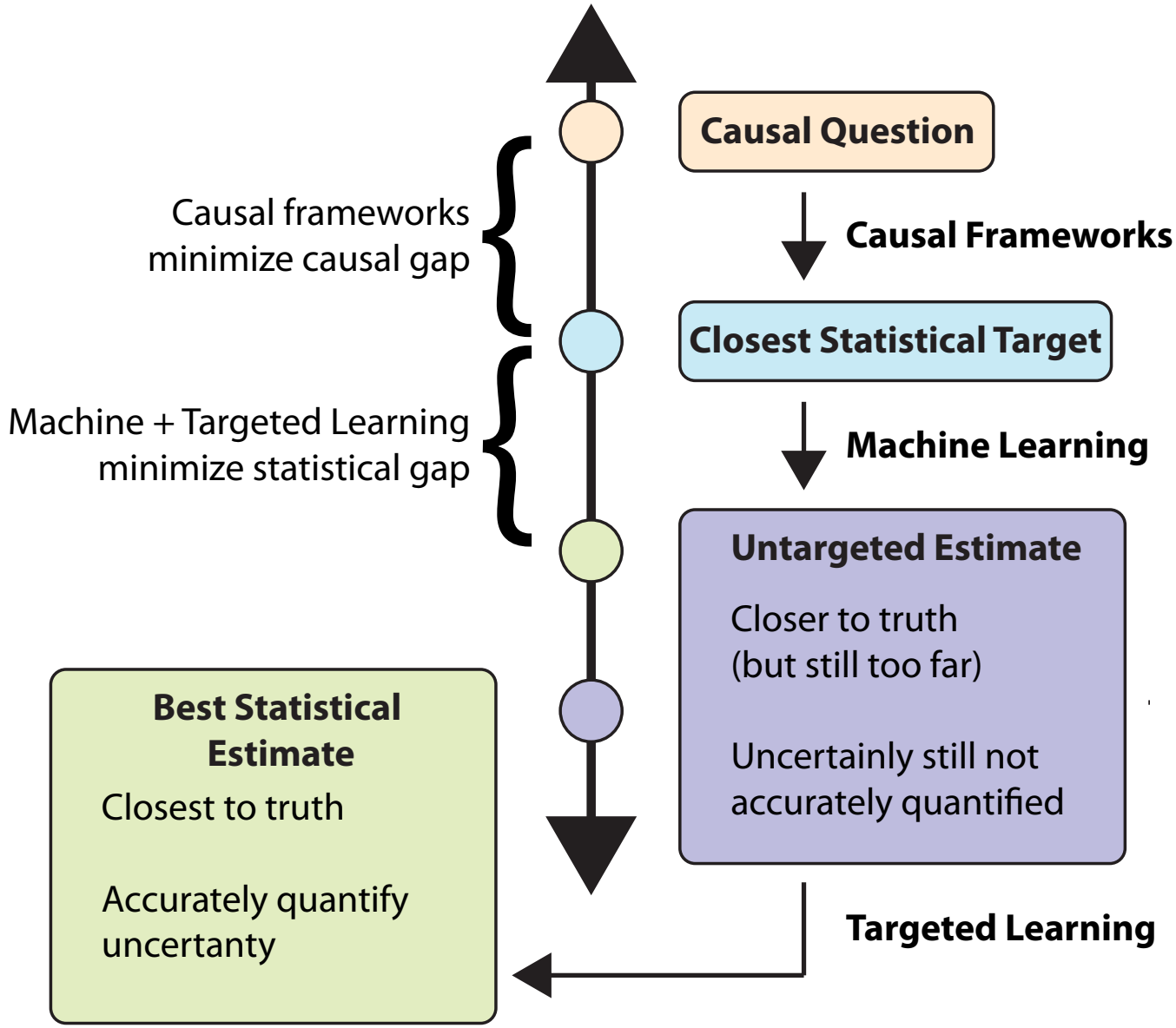
Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

The Statistical Crisis in Science

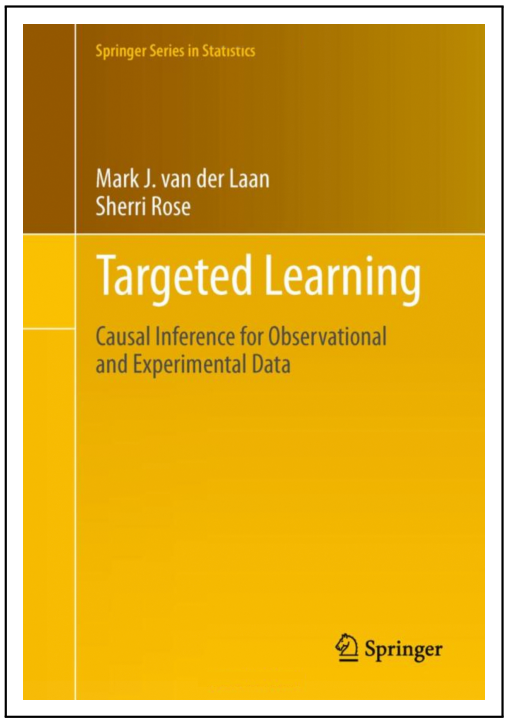
Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Andrew Gelman and Eric Loken

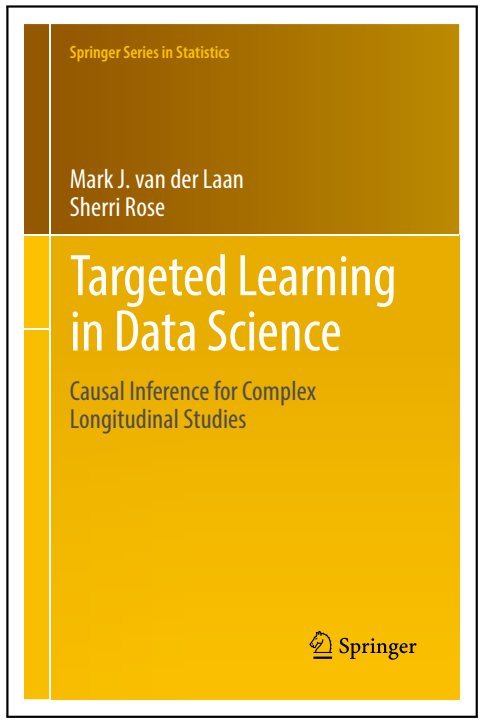
Targeted Learning for answering statistical and causal questions with confidence intervals



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`

Better clinical decisions from observational data

Statistics in Medicine

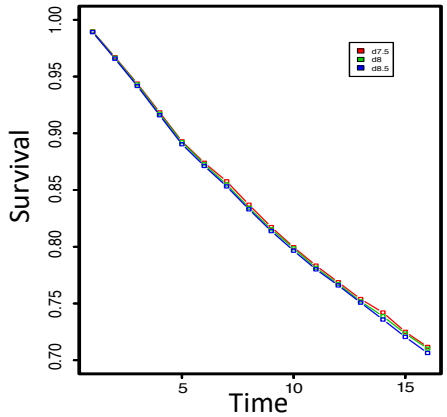
Research Article

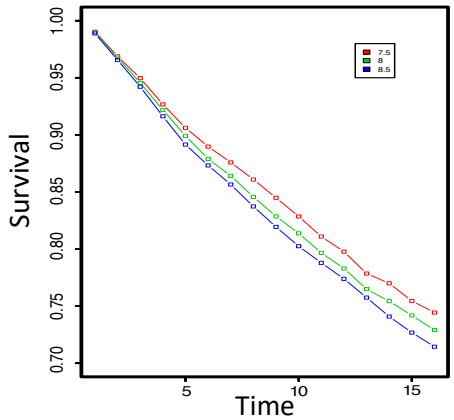
Received 24 May 2013, Accepted 5 January 2014 Published online 17 February 2014 in Wiley Online Library

(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

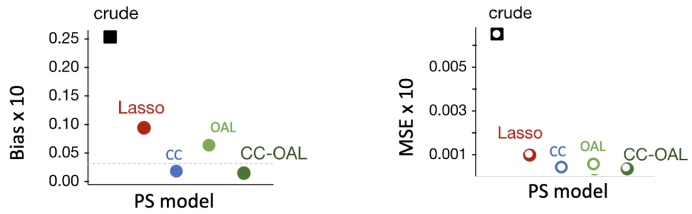
Targeted Learning: More aggressive intensification protocols result in better outcomes

FDA Sentinel Innovation Center

- **Safety evaluation with high dimensional data:**
  Wyss et al. (2024), AJE, Targeted Learning with an Undersmoothed Lasso Propensity Score Model for Large Scale Covariate Adjustment i Healthcare Database Studies.

- **Subset calibration/two-stage designs:**
  -Ongoing project evaluating methods such as the two-stage design TMLE for study designs that involve a subset of subjects with carefully curated confounders and or outcomes, and a remaining set of subjects.
  -This is a common type of design to obtain desired causal identification from RWD while still gaining efficiency from the less curated data set.

Plasmode study results



Collaborative control greatly reduced bias and improved MSE
- Less regularization captured more relevant confounder information in PS

These projects involve multi-author working groups with FDA/Pharma/Academics/Kaiser Permanente.

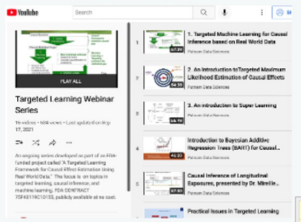
The Sentinel Innovation Center is funded by the FDA through the Department of Health and Human Services (HHS) Task order 75F40119D10037.

Comparative Effectiveness: Targeted Learning FDA Demonstration Project

Resulted in various collaborative relations with FDA statisticians

TL for RWE multi-year FDA-funded project

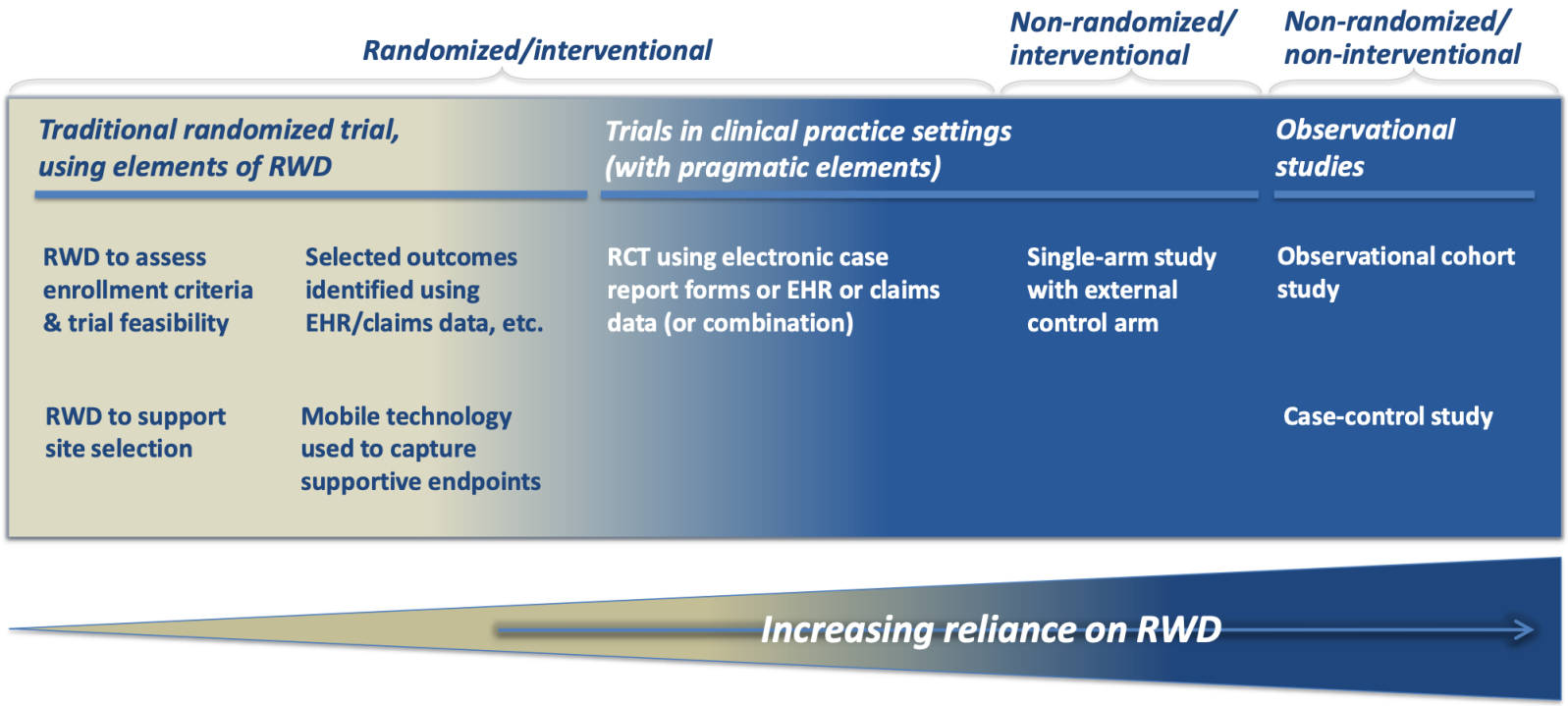
TL on YouTube



Publications with FDA co-authors

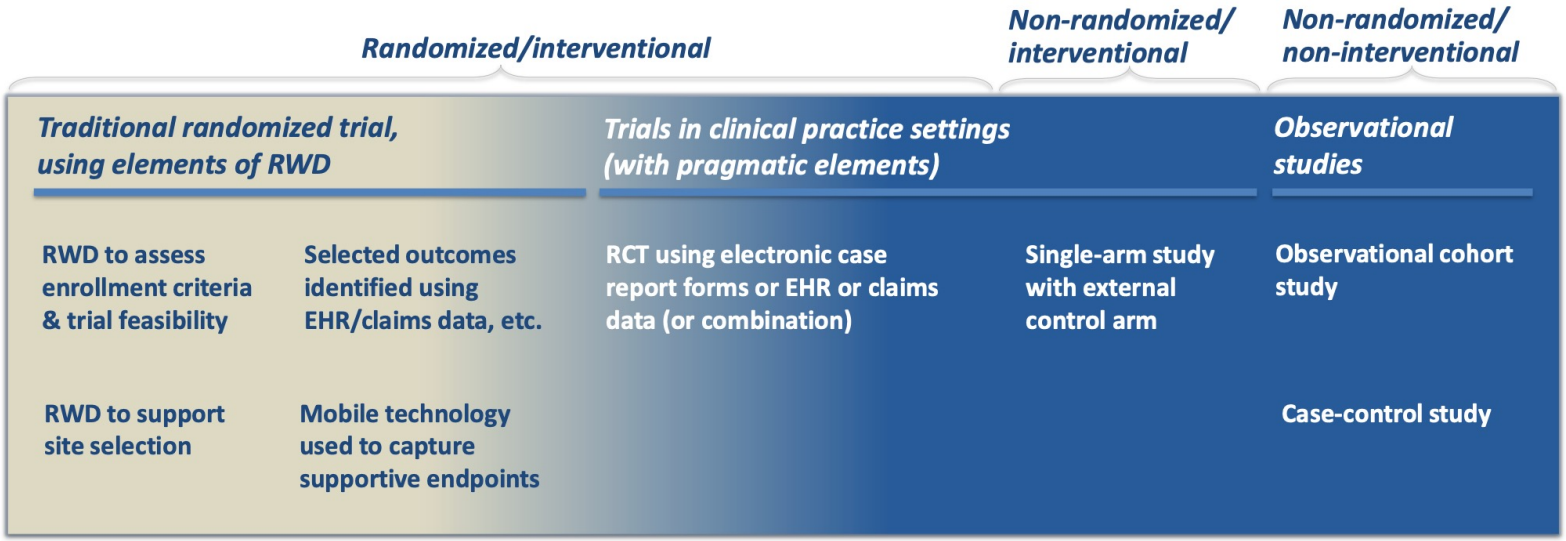
Over 250 FDA short course attendees

Statistical challenges with RWD



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

Statistical challenges with RWD



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

The roadmap for targeted learning from data

STEP 1:
DESCRIBE
EXPERIMENT

STEP 2:
SPECIFY STATISTICAL
MODEL

STEP 3:
DEFINE STATISTICAL
QUERY

STEP 4:
CONSTRUCT
ESTIMATOR

STEP 5:
OBTAIN
INFERENCE

STEP 6:
MAKE SUBSTANTIVE
CONCLUSION

Targeted Maximum Likelihood Estimation (TMLE)

STEP 1:
DESCRIBE
EXPERIMENT

STEP 2:
SPECIFY
STATISTICAL MODEL

STEP 3:
DEFINE STATISTICAL
QUERY

STEP 4:
CONSTRUCT
ESTIMATOR

STEP 5:
OBTAIN
INFERENCE

STEP 6:
MAKE SUBSTANTIVE
CONCLUSION

TMLE

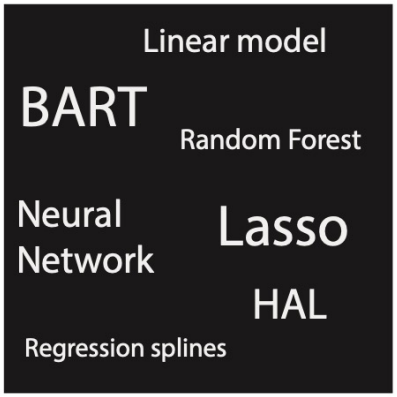
1. Initial estimation of $E[Y|A,W]$ with super (machine) learning
2. Updating initial estimate to acheive optimal bias-variance trade-off for ψ_{stat}

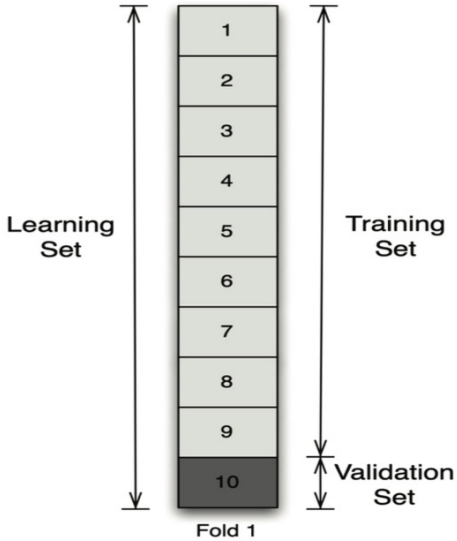
TMLE estimates are optimal:

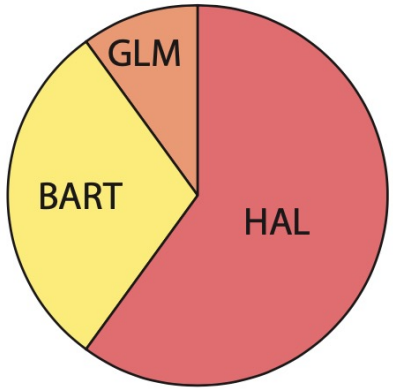
plug-in, efficient, unbiased, finite sample robust

TMLE Step 1: Super learner

Cross-validated performance of learners + ensembles



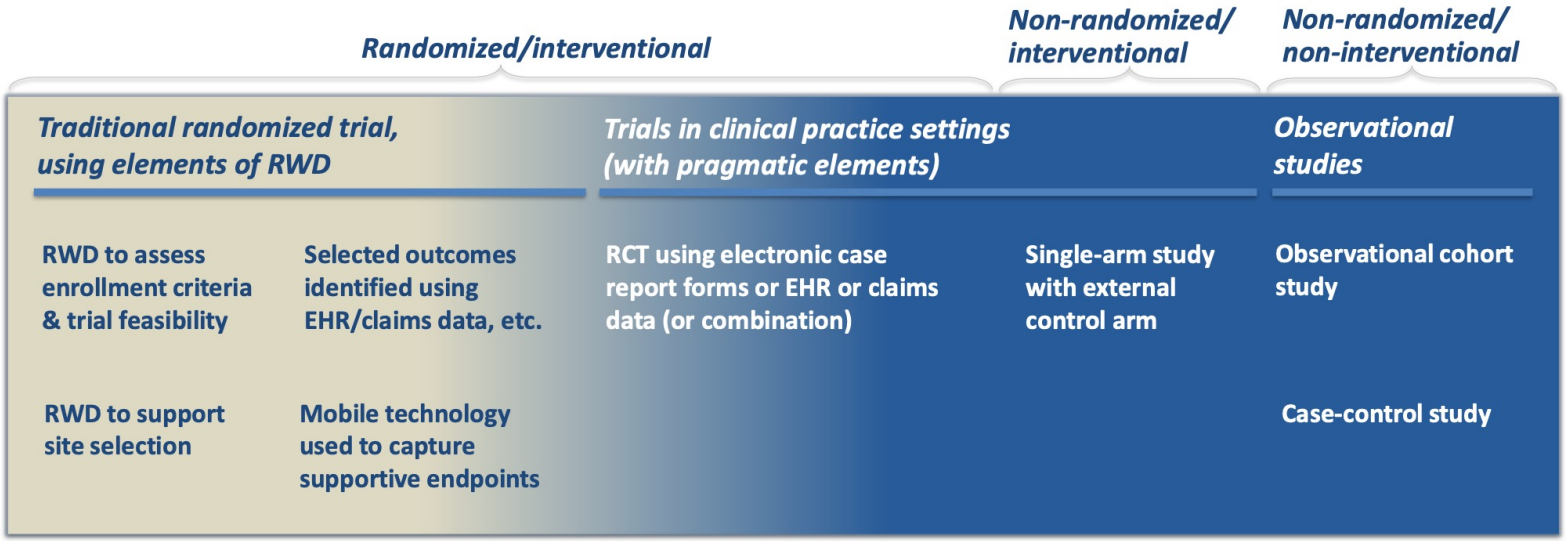




Hugely advantageous when coupled with NLP-derived covariates with EHR

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

Targeted Learning with RWD



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

A typical rare disease RCT (FDA/TLRev/UCBerkeley collaboration)

- On each subject we observe baseline history W; a binary randomized treatment A; at multiple visit times an outcome process $Y(t)$ at $t = 1, \ldots, \tau$.
- The outcome $Y(t)$ has multiple components, $Y_k(t)$, $k = 1, \ldots, K$.
- Sample size small, e.g. $n = 50$, $g_0(1|W) = P_0(A = 1) = 2/3$.
- One defines **some** composite outcome such as a sum of scores $\bar{Y} = \sum_{k=1}^{K} Y(k)$ and define the causal estimand $\Psi(P_0)$ as the ATE on $\bar{Y}$.
- A TMLE involves super-learning fit $\bar{Q}_n$ of $E_0(Y \mid W, A)$; a targeted update $\bar{Q}_n^*$ involving true PS $g_0(1|W)$, and plug-in estimator $1/n \sum_i \{\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)\}$.
- Such a TMLE is unbiased (due to DR) and typically heavily outperforms a simple unadjusted estimator of the ATE.

Simulations imitating real RCT from Zevra

Simulation Setup:

- Sample size: $n = 50$.
- 5 baseline covariates: 2 continuous, 2 binary, 1 count.
- 2:1 randomization (treatment:control).
- Outcome: Sum of multi-domain assessment scores (0-20). True outcome model is nonlinear with 2 prognostic covariates (often not known due to poor natural history understanding, necessitates the use of SL).

Results (True ATE = -1.486; 500 simulations):

Method	Bias	SE	MSE	Coverage	Power
Unadjusted	-0.096	1.829	3.356	0.946	0.118
ANCOVA	0.019	0.424	0.180	0.940	0.910
TMLE+SL	-0.016	0.369	0.137	0.940	0.988

TMLE+SL achieves efficiency gains over fixed adjustment methods (e.g., ANCOVA) while maintaining valid inference.

Right-censoring of outcome process in RCT

- Typically, a percentage of the subjects are right-censored.
- Forward imputation of outcome after drop-out is problematic.
- A common method is mixed linear repeated measures models (MLRM).
- MLRM remains consistent under informative right-censoring if the model for $E(Y(t)|W,A) = m_\beta(t,W,A)$ is correctly specified. However, in general, it is inconsistent due to informative drop-out.
- Instead, one might use ltmle() to evaluate the ATE in the world in which subjects are uncensored till end-point.
- Pros: Allows informative drop-out; utilizes SL to gain efficiency.

Simulations of ltmle() versus MLRM

Simulation Setup:

- Sample size, baseline covariates and treatment assignment same as in the point treatment setting.
- Longitudinal data with 4 time points
- Informative censoring depends on treatment and last observed outcome.
- True outcome process remains nonlinear (MLRM is misspecified).

Results (True ATE = -1.499; 500 simulations):

Method	Bias	SE	MSE	Coverage	Power
MLRM	-0.137	0.526	0.296	0.896	0.874
L-TMLE+SL	0.045	0.469	0.222	0.950	0.864

MLRM yields biased estimates and invalid inference due to mean model misspecification. L-TMLE remains unbiased and maintains nominal coverage (efficiency gain due to SL).

Multiple testing Challenge

- One is concerned about just defining some sum score outcome or other choice.
- We might lose power by that choice.
- One could define an ATE $\Psi_k(P_0)$ for each outcome $Y(k)$, $k = 1, \ldots, K$.
- Compute TMLE $\psi_n^*(k) = 1/n \sum_i \{\bar{Q}_{k,n}^*(1, W_i) - \bar{Q}_{k,n}^*(0, W_i)\}$.
- A vector TMLE $(\psi_n^*(k) : k = 1, \ldots, K)$ satisfies $n^{1/2}(\psi_n^* - \psi_0)/\sigma_n \Rightarrow_d N(0, \Sigma_0)$ with Σ_0 being correlation matrix of the vector influence curve $(D^*_{\Psi_k(),P_0}/\sigma(k) : k)$ of the TMLE.

Multivariate Normal Null Distribution from Influence Curve TMLE (Dudoit, vdL MT book)

- $N(0, \Sigma_0)$ represents the right null distribution for multiple testing of $H_0(k) : \Psi_k(P_0) = 0$, $k = 1, \ldots, K$. One sets the cut-offs of the t-statistics $n^{1/2}(\psi_n^*(k) - \psi_0(k))/\sigma_n(k)$, $k = 1, \ldots, K$, so that the FWE or any other generalized type I error is controlled at level 0.05 under sampling the t-statistic vector from $N(0, \Sigma_0)$.
- Similarly, one can use this null distribution $N(0, \Sigma_0)$ in a step-down multiple testing procedure.
- Using quantile-quantile function one can transform the marginal distributions of $N(0, \rho_0)$ into marginal distributions controlled by user such as a permutation distribution.
- Even though this is much more powerful than Bonferroni, small sample sizes generally imply lack of power for any multiple testing procedure.

Using Max-t statistic to test overall null $\Psi(P_0) = 0$

- To avoid multiple testing, one aims to test the overall null $H_0 : \psi_0(k) = 0,\ k = 1, \ldots, K$.
- In short: $H_0 : \psi_0 = 0$.
- One could do that with the max-t statistic applied to standardized TMLE and setting cut-off to control type-I error under the $N(0, \Sigma_0)$-distribution of the t-statistic.
- Much better than Bonferoni, but still price to pay in power!

Define α-specific ATE and test of its null $H_0(\alpha)$

- Let $\bar{Y}(\alpha) \equiv \sum_{k=1}^{K} \alpha(k) Y(k)$.
- Define ATE

$$\psi_0(\alpha) = E_0\{E_0(\bar{Y}(\alpha) \mid A = 1, W) - E_0(\bar{Y}(\alpha) \mid A = 0, W)\}.$$

- The TMLE of $\psi_0(\alpha)$ using an inconsistent estimator $\bar{Q}^*_{\alpha,n} = \sum_{k=1}^{K} \alpha(k) \bar{Q}^*_{k,n}$ of $\bar{Q}_{0,\alpha} = E_0(\bar{Y}(\alpha) \mid A, W) = \sum_k \alpha(k) E_0(Y(k) \mid W, A)$ is asymptotically linear with influence curve $\sum_{k=1}^{K} \alpha(k) D^*_{\Psi_k(), \bar{Q}, g_0}$.
- $\bar{Q} = (\bar{Q}_k : k)$ represents limit of outcome regressions.
- The asymptotic variance of the standardized TMLE $\psi^*_n(\alpha)$:

$$\begin{aligned} \sigma^2_{\alpha,0} &= \sigma^2_\alpha(\bar{Q}, g_0, P_0) = \alpha^\top \Sigma_0 \alpha \\ \Sigma_0(k_1, k_2) &= P_0 D^*_{\Psi_{k_1}(), \bar{Q}, g_0} D^*_{\Psi_{k_2}(), \bar{Q}, g_0}. \end{aligned}$$

Oracle test

- The power of the test $n^{1/2}\psi_n^*(\alpha)/\sigma_{\alpha,0} > 1.65$ at $\psi_0(\alpha)$ is given by

$$\bar{\Phi}(1.65 - n^{1/2}\psi_0(\alpha)/\sigma_{\alpha,0})$$

with $\bar{\Phi}(x) = P(N(0,1) > x)$.

- Define

$$\alpha_0 = \arg\max_{\alpha} \psi_0(\alpha)/\sigma_{\alpha,0}.$$

- Define oracle shift $\theta_{\alpha,0} \equiv \psi_0(\alpha)/\sigma_{\alpha,0}$.
- Then $H_0(\alpha_0) : \psi_0(\alpha_0) = 0$ implies the most powerful test among all tests.
- We have

$$\alpha_0 = \Sigma_0^{-1}(\psi_0),$$

with $\psi_0 = (\psi_0(k) : k = 1, \ldots, K)$.

Estimated oracle test statistic

- Suppose we estimate α_0 with $\alpha_n = \Sigma_n^{-1}\psi_n^*$.
- The t-statistic for $H_0(\alpha_n)$ is given by $t_n = n^{1/2}\alpha_n^\top \psi_n^*/\sigma_{\alpha_n,n}$ which can be written as

$$t_n = n^{1/2}\left(\psi_n^{*,\top}\Sigma_n^{-1}\psi_n^*\right)^{1/2}.$$

- Thus the estimate of the oracle t-statistic yields a TMLE-based Hotelling Chi-square statistic:

$$t_n^2 = n\left(\psi_n^{*,\top}\Sigma_n^{-1}\psi_n^*\right)^{1/2} \sim_{H_0} X_K^2.$$

- This suggests that latter test is highly powerful!

Data adaptive target parameter using sample splitting

- Use V-fold sample splitting with empirical measures $P_{n,v}, P^1_{n,v}$, $v = 1, \ldots, V$, for training and validation sample across splits.
- Compute $\hat{\alpha}(P_{n,v})$ of oracle α_0 based on training sample.
- Consider data adaptive target parameter

$$\theta_{n,v,0} = \alpha^\top_{n,v}\psi_0/\sigma_{\alpha_{n,v}}(\bar{Q}, g_0, P_0).$$

- Due to $\frac{d}{d\alpha_0}\alpha_0^\top\psi_0/\sigma_{\alpha,0} = 0$ we have that

$$\theta_{n,v,0} \approx \alpha_0^\top\psi_0/\sigma_{\alpha_0,0}$$

in first order!

Cross-fitted TMLE of data adaptive target parameter

- Compute a cross-fitted TMLE $\theta^*_{n,v,n} \equiv \alpha^\top_{n,v}\psi^*_{n,v}/\sigma_{\alpha_{n,v},n}$ of this $\theta_{n,v,0}$.
- Here the TMLE $\psi^*_{n,v}$ gets its initial estimator still from training sample but targeting is carried out on validation sample $P^1_{n,v}$.
- The latter targeting step can be pooled across sample splits v.
- One can estimate $\sigma_{\alpha_{n,v},0}$ based on whole sample, no sample splitting needed.
- The CV-TMLE of $\theta_{\hat{\alpha},0} = 1/V \sum_v \theta_{n,v,0}$ is defined as the average:

$$\theta_{\hat{\alpha},cv-tmle,n} = \frac{1}{V}\sum_{v=1}^{V} \theta^*_{n,v,n}.$$

Asymptotics of CV-TMLE, Inference for fixed oracle parameter as well

- **Suppose that the overall null does not hold.**
- Then α_n converges to the oracle choice α_0.
- One can prove (without Donsker class condition) that

$$\theta_{\hat{\alpha},cv-tmle,n} - 1/V \sum_{v=1}^{V} \theta_{\alpha_{n,v},0} = P_n D^*_{\hat{\theta}_{\alpha_0}(),P_0} + o_P(n^{-1/2}),$$

  for influence curve of $\hat{\theta}_{\alpha_0}(P_n)$ treating α_0 as fixed/known.
- Thus, it provides inference for $1/V \sum_{v=1}^{V} \alpha_{n,v}^\top \psi_0 / \sigma_{\alpha_{n,v},0}$ (confidence interval and test).
- Moreover, under $o_P(n^{-1/4})$-consistency of $\hat{\alpha}(P_n)$ (we have $n^{-1/2}$) yields that it also yields inference for the fixed parameter $\alpha_0^\top \psi_0 / \sigma_{\alpha_0,0}$.
- It provides a test of $H_0 : \alpha_0^\top \psi_0 = 0$ and thus $H_0 : \psi_0(k) = 0$ for all $k = 1, \ldots, K$, as long as the overall null H_0 is not true.
- Simulations (Tianyue) show nice power.

Challenge under overall null of no treatment effect

- Under the overall null, the oracle choice α_0 is not unique/defined.
- Therefore, $\hat{\alpha}(P_n)$ will stay random as sample size grows.
- We can still get an expansion for CV-TMLE but now in terms of $1/V \sum_v P^1_{n,v} D^*_{\hat{\theta}_{\alpha_{n,v}}(),P_0}$, with a random index $\alpha_{n,v}$, causing lack of normality.
- Simulations show lack of normality but only slightly anti-conservative type-I error control.
- We can use a variance stabilized average over v as our test statistic (a la Alex Luedtke online stabilized one-step estimator!):

$$T_n \equiv n^{1/2} 1/V \sum_v \sigma^{-1}_{\alpha_{n,v},n} \Psi_{\alpha_{n,v}}(P^*_{n,v}).$$

- To be continued!

THANK YOU!