# Neyman Meets Causal Machine Learning

Kosuke Imai

Harvard University

October 18, 2023

Statistics Seminar
Department of Statistics
University of Wisconsin, Madison

Joint work with Michael Lingzhi Li (Harvard Business School)

# Motivation and Overview

- 100th anniversary of Jerzy Neyman's dissertation
  1. potential outcomes notation
  2. randomization inference for the average treatment effect
- Rise of causal machine learning (causal ML)
  1. heterogeneous treatment effects
  2. individualized treatment rules
- Experimental evaluation of causal ML under Neyman's framework
  1. causal ML algorithms may not work well in practice
  2. assumption-free uncertainty quantification is essential
- Today's talk will show how to experimentally evaluate:
  1. individualized treatment rules derived by causal ML
  2. heterogeneous treatment effects discovered by causal ML
  3. exceptional responders identified by causal ML

# Neyman's Repeated Sampling Framework

- Notation: $n$ experimental units
  1. $T_i \in \{0, 1\}$: binary treatment
  2. $Y_i(t)$ where $t \in \{0, 1\}$: potential outcomes
  3. $Y_i = Y_i(T_i)$: observed outcome

- Assumptions:
  1. no interference between units: $Y_i(T_1 = t_1, \ldots, T_n = t_n) = Y_i(T_i = t_i)$
  2. randomization of treatment assignment: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i$
  3. random sampling of units: $\{Y_i(1), Y_i(0)\} \overset{\text{i.i.d.}}{\sim} \mathcal{P}$

- Causal estimand and estimator
  1. average treatment effect (ATE): $\tau = \mathbb{E}(Y_i(1) - Y_i(0))$
  2. difference-in-means estimator: $\hat{\tau} = \frac{1}{n_1} \sum_{i=1}^{n} Y_i T_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) Y_i$

- Finite sample results
  1. unbiasedness: $\mathbb{E}(\hat{\tau}) = \tau$
  2. variance: $\mathbb{V}(\hat{\tau}) = \frac{\mathbb{V}(Y_i(1))}{n_1} + \frac{\mathbb{V}(Y_i(0))}{n_0}$

# 1. Individualized Treatment Rules

# Experimental Evaluation of Individualized Treatment Rules

- Consider a fixed (for now) individualized treatment rule (ITR):

$$f(X_i) \in \{0, 1\}$$

  where $X_i$ is a set of pre-treatment covariates
  - ITR is obtained from an external dataset (e.g., sample splitting)
  - no assumption about ITR (e.g., any causal ML, heuristic rule)

- Evaluation metric examples:
  1. Population average value (PAV)

  $$\lambda_f \;=\; \mathbb{E}\{Y_i(f(X_i))\}$$

  2. Population average prescriptive effect (PAPE)

  $$\gamma_f \;=\; \mathbb{E}\{Y_i(f(X_i)) - pY_i(1) - (1-p)Y_i(0)\}$$

     where $p = \Pr(f(\mathrm{X}_i) = 1)$ is the proportion treated under the ITR
  3. Difference in PAV between two ITRs

# Neyman's Inference for the Population Average Value

- A natural estimator:

$$\hat{\lambda}_f = \frac{1}{n_1} \underbrace{\sum_{i=1}^{n} Y_i f(X_i) T_i}_{\substack{\text{treated units who should} \\ \text{be treated}}} + \frac{1}{n_0} \underbrace{\sum_{i=1}^{n} Y_i (1 - f(X_i))(1 - T_i)}_{\substack{\text{untreated units who should} \\ \text{not be treated}}},$$

- Unbiasedness: $\mathbb{E}(\hat{\lambda}_f) = \lambda_f$
- Variance:

$$\mathbb{V}(\hat{\lambda}_f) = \frac{\mathbb{V}\{f(X_i) Y_i(1)\}}{n_1} + \frac{\mathbb{V}\{(1 - f(X_i)) Y_i(0)\}}{n_0}$$

  where all observations are used to estimate the variance

- Similar results for the PAPE with a negligible finite-sample bias due to estimation of the proportion treated $p$

# Using the Same Data for Learning and Evaluation

- Cross-fitting procedure:
  1. randomly split the data into $K$ folds: $Z_1, \ldots, Z_K$
  2. learn an ITR using $K-1$ folds: $\hat{f}_{-k}$
  3. evaluate it with the held-out set: $\hat{\lambda}_{\hat{f}_{-k}}(Z_k)$
  4. repeat the process for each $k$ and compute an average
- Additional assumption: random splitting
- ML algorithm:

$$F : \mathcal{Z} \longrightarrow \mathcal{F}$$

where $Z^{\text{train}} \in \mathcal{Z}$ and $\hat{f} = F(Z^{\text{train}}) \in \mathcal{F}$

- Estimand and unbiased estimator:

$$\lambda_F = \underbrace{\mathbb{E}\{Y_i(\hat{f}_{Z^{\text{train}}}(X_i))\}}_{\text{average performance of } F}, \quad \hat{\lambda}_F = \frac{1}{K}\sum_{k=1}^{K} \hat{\lambda}_{\hat{f}_{-k}}(Z_k)$$

- Unbiasedness: $\mathbb{E}(\hat{\lambda}_F) = \lambda_F$

# Finite-sample Variance with Cross-fitting

- Correlation due to the overlap between training and evaluation data:

$$\mathbb{V}(\hat{\lambda}_F) \;=\; \frac{\mathbb{V}(\hat{\lambda}_{\hat{f}_{-k}}(Z_k))}{K} + \frac{K-1}{K}\mathrm{Cov}(\hat{\lambda}_{\hat{f}_{-k}}(Z_k), \hat{\lambda}_{\hat{f}_{-k'}}(Z_{k'}))$$

- Useful lemma about cross-validation statistics (Nadeau and Bengio 2003):

$$\mathrm{Cov}(\hat{\lambda}_{\hat{f}_{-k}}(Z_k), \hat{\lambda}_{\hat{f}_{-k'}}(Z_{k'})) \;=\; \mathbb{V}(\hat{\lambda}_{\hat{f}_{-k}}(Z_k)) - \mathbb{E}(S_F^2)$$

where $S_F^2$ is the sample variance of $\hat{\lambda}_{\hat{f}_{-k}}(Z_k)$ across $K$ folds

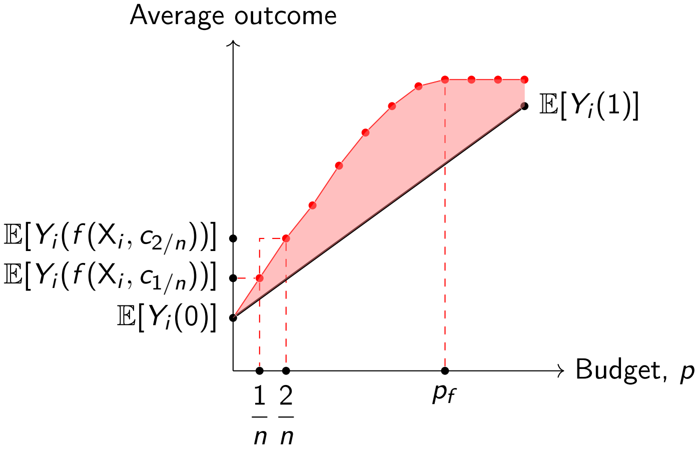
- Simplifying the expression gives:

$$\mathbb{V}(\hat{\lambda}_F) \;=\; \underbrace{\frac{\mathbb{V}\{\hat{f}_{-k}Y_i(1)\}}{n_1/K} + \frac{\mathbb{V}\{(1-\hat{f}_{-k}(X_i))Y_i(0)\}}{n_0/K}}_{\text{variance for a fixed ITR}} - \underbrace{\frac{K-1}{K}\mathbb{E}(S_F^2)}_{\substack{\text{efficiency gain} \\ \text{due to cross-fitting}}}$$

$$+\mathbb{E}\left\{\mathrm{Cov}(\hat{f}_{-k}(X_i), \hat{f}_{-k}(X_j) \mid X_i, X_j)\tau_i\tau_j\right\} \;\geq\; \mathbb{E}(S_F^2)$$

where $i \neq j$ and $\tau_i = Y_i(1) - Y_i(0)$ is the individual treatment effect

# Area Under Prescriptive Effect Curve (AUPEC)




- Measure of performance across different budget constraints
- Inference is possible with or without cross-fitting
- Normalized AUPEC = average percentage gain using an ITR over the randomized treatment rule across a range of budget contraints

# Simulations

- Atlantic Causal Inference Conference data analysis challenge
- Data generating process
  - 8 covariates from the Infant Health and Development Program (originally, 58 covariates and 4,302 observations)
  - population distribution = original empirical distribution
  - highly nonlinear model
- 5-fold cross fitting based on LASSO
- std. dev. for $n = 500$ is roughly half of the fixed $n = 100$ case

| | $n = 100$ | | | $n = 500$ | | | $n = 2000$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Estimator | cov. | bias | s.d. | cov. | bias | s.d. | cov. | bias | s.d. |
| **Small effect** | | | | | | | | | |
| PAV | 96.9 | −0.007 | 0.261 | 96.5 | −0.003 | 0.125 | 97.3 | 0.001 | 0.062 |
| PAPE | 93.6 | −0.000 | 0.171 | 93.0 | 0.000 | 0.093 | 95.3 | 0.001 | 0.041 |
| **Large effect** | | | | | | | | | |
| PAV | 96.9 | −0.007 | 0.261 | 96.5 | −0.003 | 0.125 | 97.3 | 0.001 | 0.062 |
| PAPE | 93.6 | −0.000 | 0.171 | 93.0 | 0.000 | 0.093 | 95.3 | 0.001 | 0.041 |

# Application to the STAR Experiment

- Experiment involving 7,000 students across 79 schools
- Randomized treatments (kindergarden):
  1. $T_i = 1$: small class (13–17 students)
  2. $T_i = 0$: regular class (22–25)
- Outcome: SAT scores
- 10 covariates: 4 demographic and 6 school characteristics
- Sample size: $n = 1911$, 5-fold cross-fitting

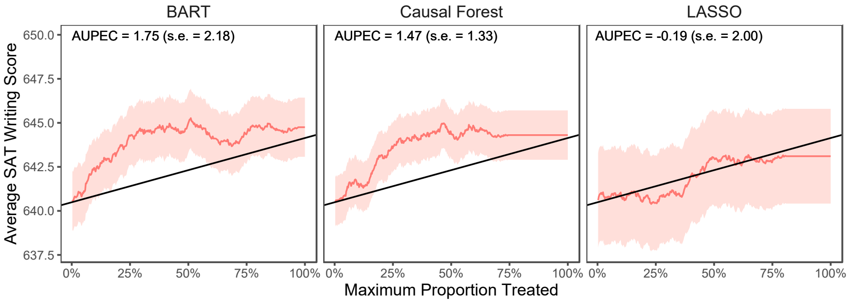
- Estimated average treatment effects:
  - SAT reading: 6.78 (s.e.=1.71)
  - SAT math: 5.78 (s.e.=1.80)
  - SAT writing:3.65 (s.e.=1.63)

# Results

- ITR performance via PAPE

| | BART | | | Causal Forest | | | LASSO | | |
|---|---|---|---|---|---|---|---|---|---|
| | est. | s.e. | treated | est. | s.e. | treated | est. | s.e. | treated |
| Reading | 0.19 | 0.37 | 99.3% | 0.31 | 0.77 | 86.6% | 0.32 | 0.53 | 87.6% |
| Math | 0.92 | 0.75 | 84.7 | 2.29 | 0.80 | 79.1 | 1.52 | 1.60 | 75.2 |
| Writing | 1.12 | 0.86 | 88.0 | 1.43 | 0.71 | 67.4 | 0.05 | 1.37 | 74.8 |

- AUPEC



# 2. Heterogeneous Treatment Effects

# Evaluation of Heterogeneous Treatment Effects

- How can we make statistical inference for heterogeneous treatment effects discovered by a generic ML algorithm?
- Conditional Average Treatment Effect (CATE):

$$\tau(x) \;=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = x)$$

- CATE estimation based on ML algorithm

$$f : \mathcal{X} \longrightarrow \mathcal{S} \subset \mathbb{R}$$

- Sorted Group Average Treatment Effect (GATES; Chernozhukov et al. 2019)

$$\tau_k \;=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid p_{k-1} \leq S_i = f(X_i) < p_k)$$

for $k = 1, 2, \ldots, K$ where $p_k$ is a cutoff ($p_0 = -\infty$, $p_K = \infty$)

# GATES Estimation as ITR Evaluation

- A natural GATES estimator:

$$\hat{\tau}_k = \frac{K}{n_1}\sum_{i=1}^{n} Y_i T_i \hat{g}_k(X_i) - \frac{K}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)\hat{g}_k(X_i),$$

  where $\hat{g}_k(\mathrm{X}_i) = 1\{S_i \geq \hat{p}_k(s)\} - 1\{S_i \geq \hat{p}_{k-1}\}$

- Rewrite $\hat{\tau}_k$:

$$\hat{\tau}_k = K\Bigg\{\underbrace{\frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i \hat{g}_k(X_i) + \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)(1-\hat{g}_k(X_i))}_{\text{estimated PAV of } \hat{g}_k} - \underbrace{\frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)}_{\text{PAV of treat-no-one policy}}\Bigg\}$$

- We can directly apply our previous results
- Inference for GATES under cross-fitting is possible too
- Statistical hypothesis tests of treatment effect heterogeneity

# Empirical Application

- National Supported Work Demonstration Program (LaLonde 1986)
- Temporary employment program to help disadvantaged workers by giving them a guaranteed job for 9 to 18 months

- Data
  - sample size: $n_1 = 297$ and $n_0 = 425$
  - outcome: annualized earnings in 1978 (36 months after the program)
  - 7 pre-treatment covariates: demographics and prior earnings

- Setup
  - ML algorithms: Causal Forest, BART, and LASSO
  - Sample-splitting: 2/3 of the data as training data
  - Cross-fitting: 3 folds

# GATES Estimates (in 1,000 US Dollars)

| | $\hat{\tau}_1$ | $\hat{\tau}_2$ | $\hat{\tau}_3$ | $\hat{\tau}_4$ | $\hat{\tau}_5$ |
|---|---|---|---|---|---|
| **Sample-splitting** | | | | | |
| BART | 2.90 | −0.73 | −0.02 | 3.25 | 2.57 |
| | [−2.25, 8.06] | [−5.05, 3.58] | [−3.47, 3.43] | [−1.53, 8.03] | [−3.82, 8.97] |
| Causal Forest | 3.40 | 0.13 | −0.85 | −1.91 | 7.21 |
| | [−1.29, 3.40] | [−5.37, 5.63] | [−5.22, 3.52] | [−5.16, 1.34] | [1.22, 13.19] |
| LASSO | 1.86 | 2.62 | −2.07 | 1.39 | 4.17 |
| | [−3.59, 7.30] | [−1.69, 6.93] | [−5.39, 1.26] | [−2.95, 5.73] | [−2.30, 10.65] |
| **Cross-fitting** | | | | | |
| BART | 0.40 | −0.15 | −0.40 | 2.52 | 2.19 |
| | [−3.79, 4.59] | [−2.54, 2.23] | [−3.37, 2.56] | [−0.99, 6.03] | [−0.73, 5.11] |
| Causal Forest | −3.72 | 1.05 | 5.32 | −2.64 | 4.55 |
| | [−6.52, −0.93] | [−2.28, 4.37] | [2.63, 8.01] | [−5.07, −0.22] | [1.14, 7.96] |
| LASSO | 0.65 | 0.45 | −2.88 | 1.32 | 5.02 |
| | [−3.65, 4.94] | [−3.28, 4.18] | [−5.38, −0.38] | [−1.83, 4.48] | [−0.14, 10.18] |

# 3. Exceptional Responders

# Identification of Exceptional Responders

- In the GATES estimation, the cutoff $p$ is given
- Goal: provide a statistical guarantee when selecting $p$ using the data
- The problem is trivial if we had an infinite amount of data

$$p^* = \operatorname*{argmax}_{p \in [0,1]} \Psi(p) \quad \text{where } \Psi(p) \; = \; \mathbb{E}[\underbrace{Y_i(1) - Y_i(0)}_{=\psi_i} \mid F(S_i) \geq p],$$

1. sample size may not be large
2. ML estimates of CATE may be biased and noisy
3. proportion of exceptional responders may be small

- Standard method suffers from multiple testing problem:

$$\hat{p}_n \; = \; \operatorname*{argmax}_{p \in [0,1]} \widehat{\Psi}_n(p) \quad \text{where } \widehat{\Psi}_n(p) \; = \; \frac{1}{np} \sum_{i=1}^{\lfloor np \rfloor} \hat{\psi}_{[n,i]}$$

where $S_{[n,1]} \geq S_{[n,2]}, \ldots, \geq S_{[n,n]}$ and

$$\hat{\psi}_{[n,i]} = \frac{T_{[n,i]} Y_{[n,i]}}{n_1/n} - \frac{(1 - T_{[n,i]}) Y_{[n,i]}}{n_0/n}$$

# Providing a Statistical Guarantee

- (one-sided) Uniform confidence band:

$$\mathbb{P}\left(\forall p \in [0,1],\ \Psi(p) \geq \widehat{\Psi}_n(p) - C_n(p,\alpha)\right) \geq 1-\alpha.$$

- Safe identification of exceptional responders:

$$\underline{\hat{p}}_n \ = \ \operatorname*{argmax}_{p\in[0,1]} \widehat{\Psi}_n(p) - C_n(p,\alpha),$$

implying

$$\begin{aligned}\mathbb{P}\left(\Psi(p^*) \geq \widehat{\Psi}_n(\underline{\hat{p}}_n) - C_n(\underline{\hat{p}}_n,\alpha)\right) &\geq \ \mathbb{P}\left(\Psi(\underline{\hat{p}}_n) \geq \widehat{\Psi}_n(\underline{\hat{p}}_n) - C_n(\underline{\hat{p}}_n,\alpha)\right) \\ &\geq \ 1-\alpha.\end{aligned}$$

- Other data-driven selection of $p$ is possible: e.g., for a given $c$

$$\begin{aligned}\text{estimate}\ \ \underline{\hat{p}}_n(c) \ &= \ \sup\{p\in[0,1] : \widehat{\Psi}_n(p) - C_n(p,\alpha) \geq c\}, \\ \text{to target}\ \ p^*(c) \ &= \ \sup\{p\in[0,1] : \Psi(p) \geq c\}\end{aligned}$$

# Constructing Uniform Confidence Band

1. Obtain finite-sample bias bound and variance of $\widehat{\Psi}_n(p)$ using our previous result
2. Use a generalized version of Donsker's invariance principle to show: for $i = 1, 2, \ldots, n$

$$\left( \frac{\mathbb{V}(\frac{i}{n}\widehat{\Psi}_n(\frac{i}{n}))}{\mathbb{V}(\widehat{\Psi}_n(1))}, \frac{\frac{i}{n}\Psi(\frac{i}{n}) - \frac{i}{n}\widehat{\Psi}_n(\frac{i}{n})}{\sqrt{\mathbb{V}(\widehat{\Psi}_n(1))}} \right) \xrightarrow{D} (p, G(p)).$$

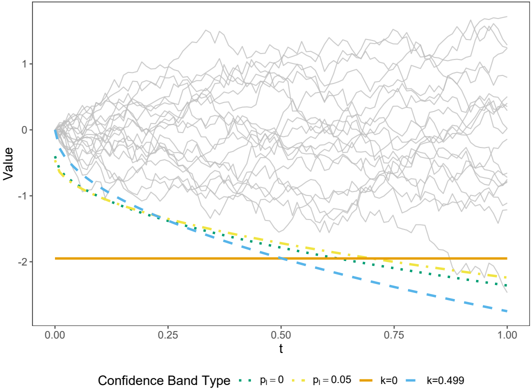
3. Show sorted individual treatment effects are non-negatively correlated

$$\mathrm{Cov}(\hat{\psi}_{[n,i]}, \hat{\psi}_{[n,j]}) \geq 0 \quad \text{for any } 1 \leq i < j \leq n$$

4. Use Slepian's Lemma to bound non-negatively correlated and normalized $p\widehat{\Psi}_n(p)$ by an appropriately scaled Wiener process
5. Approximate the confidence band by minimizing the area

$$\mathbb{P}\left( W(t) \leq \beta_0 + \beta_1\sqrt{t},\ \forall t \in [0, 1] \right) \geq 1 - \alpha$$

# Minimum-Area Confidence Band




$$\lim_{n\to\infty} \mathbb{P}\left(\forall p \in [0,1], \Psi(p) \geq \widehat{\Psi}_n(p) - \frac{\beta_0^*(\alpha)}{p}\sqrt{\mathbb{V}(\widehat{\Psi}_n(1))} - \beta_1^*(\alpha)\sqrt{\mathbb{V}(\widehat{\Psi}_n(p))}\right) \geq 1-\alpha$$

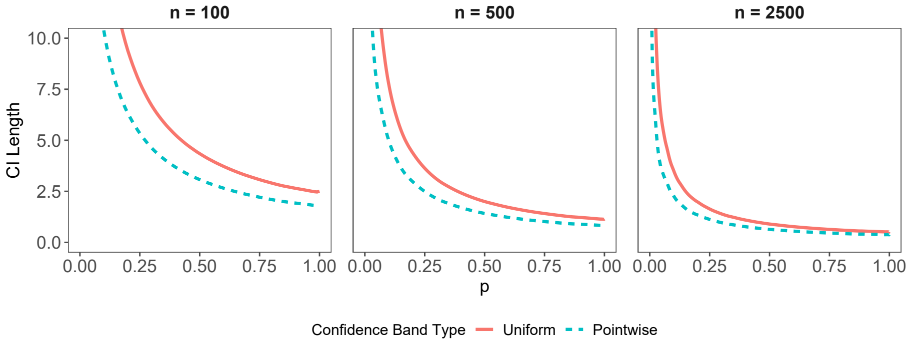
where $\{\beta_0^*(\alpha), \beta_1^*(\alpha)\}$ are the solution to:

$$\underset{\beta_0,\beta_1\in\mathbb{R}_+^2}{\text{argmin}} \int_0^1 \beta_0 + \beta_1\sqrt{t}\, dt \;\; \text{subject to} \;\; \mathbb{P}\left(W(t) \leq \beta_0 + \beta_1\sqrt{t},\ \forall t \in [0,1]\right) \geq 1-\alpha.$$

# Simulation Studies

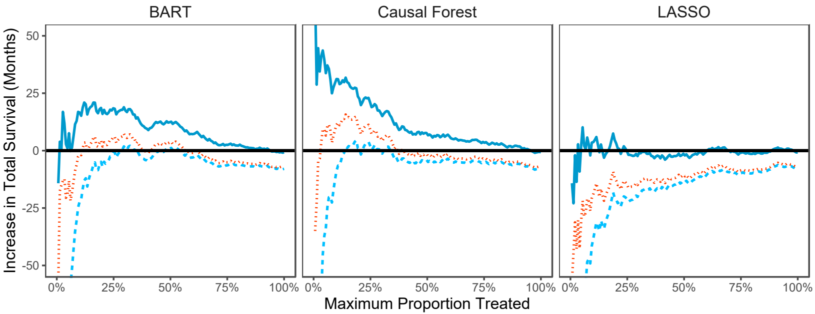
- A data generating process from the ACIC

| ML algorithm | Uniform | | | Pointwise | | |
|---|---|---|---|---|---|---|
| | $n = 100$ | $n = 500$ | $n = 2500$ | $n = 100$ | $n = 500$ | $n = 2500$ |
| BART | 96.1% | 96.0% | 95.2% | 87.2% | 76.5% | 70.3% |
| Causal Forest | 96.0% | 95.3% | 95.7% | 83.7% | 77.1% | 71.9% |
| LASSO | 95.8% | 95.6% | 95.6% | 84.1% | 76.0% | 69.8% |



# Empirical Application

- Clinical trial data on late-stage prostate cancer ($n_1 = 125$, $n_0 = 127$)
- Outcome: total survival in months, Treatment: estrogen
- Sample-split (40% train., 60% eval.), ATE estimate $-0.3$ month




| ML algorithm | Estimated proportion of exceptional responders | Estimated GATES | 90% uniform confidence band |
|---|---|---|---|
| Causal Forest | 18.8% | 27.2 | $(4.45, \infty)$ |
| BART | 32.2% | 18.1 | $(2.12, \infty)$ |
| LASSO | 91.2% | 1.35 | $(-6.26, \infty)$ |

# Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)
- Safe deployment of causal ML requires uncertainty quantification
  - Neyman's framework for experimental evaluation of HTEs and ITRs
  - No modeling assumption, Computational efficiency
  - Applicable to any complex causal ML algorithms
  - Good small sample performance
- Open source software: evalITR: Evaluating Individualized Treatment Rules at CRAN `https://CRAN.R-project.org/package=evalITR`
- More information: `https://imai.fas.harvard.edu/research/`