Experimental Evaluation of Machine Learning Algorithms for Causal Inference

Kosuke Imai

Harvard University

2021 Pacific Causal Inference Conference
September 11, 2021

Joint work with Michael Lingzhi Li (MIT)

Motivation

- Use of machine learning (ML) algorithms in experimental studies
  1. estimate heterogeneous treatment effects
  2. construct individualized treatment rules
- Software implementation of various ML algorithms is readily available
- But, do ML algorithms "work" in practice?
  - unknown theoretical properties
  - difficulty of uncertainty quantification
- We should empirically evaluate the performance of ML algorithms
  1. avoid assuming the "nice properties" of ML algorithms
  2. accurately quantify uncertainty
  3. allow for any ML algorithm
  4. applicable even when the sample size is small

Overview

- Individualized treatment rules (ITRs)
  - designed to increase efficiency of policies or treatments
  - personalized medicine, micro-targeting in business/politics
- Existing literature:
  1. development of optimal ITRs
  2. estimation of heterogeneous treatment effects
  3. extensive use of machine learning (ML) algorithms
- Goal: use a randomized experiment to *evaluate generic ITRs*
  1. Neyman's repeated sampling framework
     - randomized treatment assignment, random sampling
     - no modeling assumption or asymptotic approximation
     - extend analysis to cross-fitting regime
  2. Evaluation measures
     - shortcomings of existing metrics
     - incorporating a budget constraint
     - overall evaluation metric for general ITRs
  3. Extension to estimation of heterogeneous effects

Evaluation without a Budget Constraint

- Setup
  - Binary treatment: $T_i \in \{0, 1\}$
  - Pre-treatment covariates: $\mathbf{X} \in \mathcal{X}$
  - No interference: $Y_i(T_1 = t_1, T_2 = t_2, \ldots, T_n = t_n) = Y_i(T_i = t_i)$
  - Random sampling of units:

$$(Y_i(1), Y_i(0), \mathbf{X}_i) \overset{\text{i.i.d.}}{\sim} \mathcal{P}$$

  - Completely randomized treatment assignment:

$$\Pr(T_i = 1 \mid Y_i(1), Y_i(0), \mathbf{X}_i) = \frac{n_1}{n} \quad \text{where} \quad n_1 = \sum_{i=1}^{n} T_i$$

- Fixed (for now) ITR:

$$f : \mathcal{X} \longrightarrow \{0, 1\}$$

  - based on any ML algorithm or even a heuristic rule
  - sample splitting for experimental data, separate observational data

Neyman's Inference for the Standard Metric

- Standard metric (Population Average "Value" or PAV):

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

- A natural estimator:

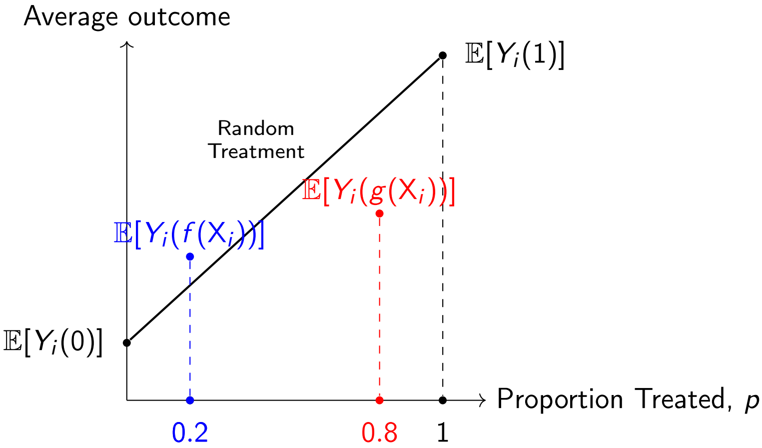
$$\hat{\lambda}_f(\mathcal{Z}) = \frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i f(X_i) + \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)(1-f(X_i)),$$

  where $\mathcal{Z} = \{X_i, T_i, Y_i\}_{i=1}^n$
- Unbiasedness: $\mathbb{E}\{\hat{\lambda}_f(\mathcal{Z})\} = \lambda_f$
- Variance:

$$\mathbb{V}\{\hat{\lambda}_f(\mathcal{Z})\} = \frac{\mathbb{E}(S_{f1}^2)}{n_1} + \frac{\mathbb{E}(S_{f0}^2)}{n_0},$$

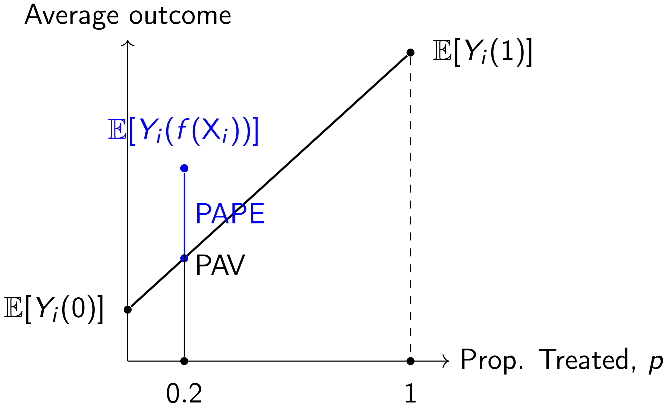
  where $S_{ft}^2 = \sum_{i=1}^n (Y_{fi}(t) - \overline{Y_f(t)})^2/(n-1)$, $Y_{fi}(t) = 1\{f(X_i) = t\}Y_i(t)$, and $\overline{Y_f(t)} = \sum_{i=1}^n Y_{fi}(t)/n$ for $t = \{0, 1\}$

A Problem of Comparing ITRs Using the PAV



- $\lambda_f < \lambda_g$: but g is performing worse than the random (i.e., non-individualized) treatment rule whereas f is not
- Need to account for the proportion treated

Accounting for the Proportion of Treated Units



- Population Average Prescriptive Effect (PAPE):

$$\tau_f = \mathbb{E}\{Y_i(f(X_i)) - p_f Y_i(1) - (1 - p_f) Y_i(0)\}$$

where $p_f = \Pr(f(X_i) = 1)$ is the proportion treated under f

Estimating the Population Average Prescriptive Effect

- An unbiased estimator of PAPE τ_f:

$$\hat{\tau}_f(\mathcal{Z}) = \frac{n}{n-1}\underbrace{\left[\frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i f(\mathrm{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)(1-f(\mathrm{X}_i))\right.}_{\text{PAV of ITR}}$$
$$- \underbrace{\left.\frac{\hat{p}_f}{n_1}\sum_{i=1}^{n} Y_i T_i - \frac{1-\hat{p}_f}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)\right]}_{\text{PAV of random treatment rule with the same treated proportion}}$$

  where $\hat{p}_f = \sum_{i=1}^{n} f(\mathrm{X}_i)/n$
- We also derive its variance, and propose its consistent estimator
- Not invariant to additive transformation: $Y_i + c$
- Solution: centering $\mathbb{E}(Y_i(1) + Y_i(0)) = 0 \rightsquigarrow$ minimum variance

Estimating and Evaluating ITRs via Cross-Fitting

- Estimate and evaluate an ITR using the same experimental data
- How should we account for both estimation uncertainty and evaluation uncertainty under the Neyman's framework?
- Setup:
  - ML algorithm
  $$F : \mathcal{Z} \longrightarrow \mathcal{F}.$$
  - K-fold cross-fitting: $\mathcal{Z} = \{\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_K\}$
  $$\hat{f}_{-k} \;=\; F(\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_{k-1}, \mathcal{Z}_{k+1}, \ldots, \mathcal{Z}_K)$$
  - Evaluation metric estimators:
  $$\hat{\lambda}_F \;=\; \frac{1}{K}\sum_{k=1}^{K} \hat{\lambda}_{\hat{f}_{-k}}(\mathcal{Z}_k), \quad \hat{\tau}_F \;=\; \frac{1}{K}\sum_{k=1}^{K} \hat{\tau}_{\hat{f}_{-k}}(\mathcal{Z}_k)$$
- Uncertainty over both evaluation data and all random sets of training data (of a fixed size) as well as treatment assignment

Causal Estimands

- Population Average Value (PAV)
  - Generalized ITR averaging over the random sampling of training data $\mathcal{Z}^{tr}$

$$\bar{f}_F(\mathbf{x}) \;=\; \mathbb{E}\{\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{x}) \mid \mathbf{X}_i = \mathbf{x}\} \;=\; \Pr(\hat{f}_{\mathcal{Z}^{tr}}(\mathbf{x}) = 1 \mid \mathbf{X}_i = \mathbf{x})$$

  - Estimand

$$\lambda_F \;=\; \mathbb{E}\left\{\bar{f}_F(\mathbf{X}_i)Y_i(1) + (1 - \bar{f}_F(\mathbf{X}_i))Y_i(0)\right\}$$

- Population Average Prescriptive Effect (PAPE)
  - Proportion treated

$$p_F \;=\; \mathbb{E}\{\bar{f}_F(\mathbf{X}_i)\}.$$

  - Estimand

$$\tau_F \;=\; \mathbb{E}\{\lambda_F - p_F Y_i(1) - (1 - p_F)Y_i(0)\}.$$

Inference under Cross-Fitting

- Under Neyman's framework, the cross-fitting estimators are unbiased, i.e., $\mathbb{E}(\hat{\lambda}_F) = \lambda_F$ and $\mathbb{E}(\hat{\tau}_F) = \tau_F$
- The variance of the PAV estimator

$$\mathbb{V}(\hat{\lambda}_F) = \underbrace{\frac{\mathbb{E}(S^2_{\hat{f}1})}{m_1} + \frac{\mathbb{E}(S^2_{\hat{f}0})}{m_0}}_{\text{evaluation uncertainty}} + \underbrace{\mathbb{E}\left\{\mathrm{Cov}(\hat{f}_{\mathcal{Z}^{tr}}(X_i), \hat{f}_{\mathcal{Z}^{tr}}(X_j) \mid X_i, X_j)\tau_i\tau_j\right\}}_{\text{estimation uncertainty}} - \underbrace{\frac{K-1}{K}\mathbb{E}(S^2_F)}_{efficiency\ gain\ due\ to\ cross-fitting}$$

for $i \neq j$ where m_t is the size of the training set with $T_i = t$, $\tau_i = Y_i(1) - Y_i(0)$, $S^2_F = \sum_{k=1}^{K}\left\{\hat{\lambda}_{\hat{f}_{-k}}(\mathcal{Z}_k) - \overline{\hat{\lambda}_{\hat{f}_{-k}}(\mathcal{Z}_k)}\right\}^2/(K-1)$

- Analogous results for the PAPE τ_F

Evaluation with a Budget Constraint

- Policy makers often face a binding budget constraint p
- Scoring rule:

$$s : \mathcal{X} \longrightarrow \mathcal{S} \quad \text{where} \quad \mathcal{S} \subset \mathbb{R}$$

- Example: CATE $s(\mathrm{x}) = \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathrm{X}_i = \mathrm{x})$
- (Fixed) ITR with a budget constraint:

$$f(\mathrm{X}_i, c) \; = \; 1\{s(\mathrm{X}_i) > c\},$$

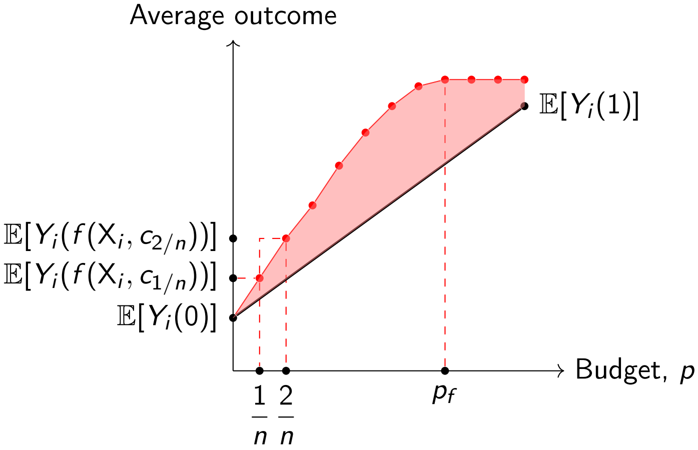
where $c_p(f) \; = \; \inf\{c \in \mathbb{R} : \Pr(f(\mathrm{X}_i, c) = 1) \leq p\}$

- PAPE under a budget constraint

$$\tau_{fp} \; = \; \mathbb{E}\{Y_i(f(\mathrm{X}_i, c_p(f))) - pY_i(1) - (1-p)Y_i(0)\}.$$

- We derive the bias (and its finite sample bound) and variance under the Neyman's framework
- Extensions: cross-fitting, diff. in PAPE between two ITRs

The Area Under Prescriptive Effect Curve (AUPEC)



- Measure of performance across different budget constraints
- We show how to do inference with and 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
  - Model

$$Y_i(t) \;=\; \mu(\mathsf{X}_i) + \tau(\mathsf{X}_i)t + \sigma(\mathsf{X}_i)\epsilon_i,$$

where $t = 0, 1$, $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, and

$$\begin{aligned}
\mu(\mathsf{x}) &= -\sin(\Phi(\pi(\mathsf{x}))) + x_{43},\\
\pi(\mathsf{x}) &= 1/[1 + \exp\{3(x_1 + x_{43} + 0.3(x_{10} - 1)) - 1\}],\\
\tau(\mathsf{x}) &= \xi(x_3 x_{24} + (x_{14} - 1) - (x_{15} - 1)),\\
\sigma(\mathsf{x}) &= 0.25\sqrt{\mathbb{V}(\mu(\mathsf{x}) + \pi(\mathsf{x})\tau(\mathsf{x}))}.
\end{aligned}$$

- Two scenarios: large vs. small treatment effects $\xi \in \{2, 1/3\}$
- Sample sizes: $n \in \{100, 500, 2,000\}$

Results I: Fixed ITR

- f: Bayesian Additive Regression Tree (BART)
- No budget constraint, 20% constraint
- g: Causal Forest
- h: LASSO

		$n = 100$			$n = 500$			$n = 2000$		
Estimator	truth	cov.	bias	s.d.	cov.	bias	s.d.	cov.	bias	s.d.
Small effect										
$\hat{\tau}_f$	0.066	94.3	0.005	0.124	96.2	0.001	0.053	95.1	0.001	0.026
$\hat{\tau}_f(c_{0.2})$	0.051	93.2	−0.002	0.109	94.4	0.001	0.046	95.2	0.002	0.021
$\widehat{\Gamma}_f$	0.053	95.3	0.001	0.106	95.1	0.001	0.045	94.8	−0.001	0.024
$\widehat{\Delta}_{0.2}(f, g)$	−0.022	94.0	0.006	0.122	95.4	0.002	0.051	96.0	0.000	0.026
$\widehat{\Delta}_{0.2}(f, h)$	−0.014	93.9	−0.001	0.131	94.9	−0.000	0.060	95.3	−0.000	0.030
Large effect										
$\hat{\tau}_f$	0.430	94.7	−0.000	0.163	95.7	0.000	0.064	94.4	-0.000	0.031
$\hat{\tau}_f(c_{0.2})$	0.356	94.7	0.004	0.159	95.7	0.002	0.072	95.8	0.000	0.035
$\widehat{\Gamma}_f$	0.363	94.3	−0.005	0.130	94.9	0.003	0.058	95.7	0.000	0.029
$\widehat{\Delta}_{0.2}(f, g)$	−0.000	96.9	0.008	0.151	97.9	−0.002	0.073	98.0	−0.000	0.026
$\widehat{\Delta}_{0.2}(f, h)$	0.000	94.7	−0.004	0.140	97.7	−0.001	0.065	96.6	0.000	0.033

Results II: Estimated ITR

- 5-fold cross fitting
- F: 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									
$\hat{\lambda}_F$	96.4	0.001	0.216	96.7	0.002	0.100	97.2	0.002	0.046
$\hat{\tau}_F$	94.6	-0.002	0.130	95.5	-0.002	0.052	94.4	-0.000	0.027
$\hat{\tau}_F(c_{0.2})$	95.4	-0.003	0.120	95.4	-0.002	0.043	96.8	0.001	0.029
$\widehat{\Gamma}_F$	98.2	0.002	0.117	96.8	-0.001	0.048	95.9	0.001	0.001
Large effect									
$\hat{\lambda}_H$	96.9	-0.007	0.261	96.5	-0.003	0.125	97.3	0.001	0.062
$\hat{\tau}_F$	93.6	-0.000	0.171	93.0	0.000	0.093	95.3	0.001	0.041
$\hat{\tau}_F(c_{0.2})$	94.8	-0.002	0.170	96.2	-0.005	0.075	95.8	0.001	0.037
$\widehat{\Gamma}_F$	98.5	0.001	0.126	98.9	0.005	0.053	99.0	0.001	0.026

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)
  3. regular class with aid
- Outcome: SAT scores
- Literature on heterogeneous treatments in labor economics
- 10 covariates
  - 4 demographics: gender, race, birth month, birth year
  - 6 school characteristics: urban/rural, enrollment size, grade range, number of students on free lunch, percentage white, number of students on school buses
- Sample size: $n = 1,911$, 5-fold cross-fitting
- Average Treatment Effects:
  - SAT reading: 6.78 (s.e.=1.71)
  - SAT math: 5.78 (s.e.=1.80)

Results I: ITR Performance

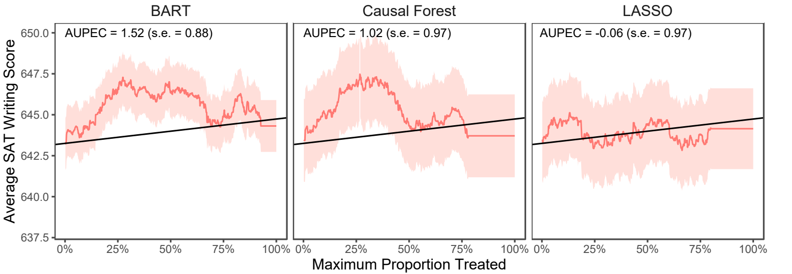
	BART			Causal Forest			LASSO		
	est.	s.e.	treated	est.	s.e.	treated	est.	s.e.	treated
Fixed ITR									
No budget constraint									
Reading	0	0	100%	−0.38	1.14	84.3%	−0.41	1.10	84.4%
Math	0.52	1.09	86.7	0.09	1.18	80.3	1.73	1.25	78.7
Writing	−0.32	0.72	92.7	−0.70	1.18	78.0	−0.30	1.26	80.0
Budget constraint									
Reading	−0.89	1.30	20	0.66	1.23	20	−1.17	1.18	20
Math	0.70	1.25	20	2.57	1.29	20	1.25	1.32	20
Writing	2.60	1.17	20	2.98	1.18	20	0.28	1.19	20
Estimated ITR									
No budget constraint									
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
Budget constraint									
Reading	1.55	1.05	20	0.40	0.69	20	−0.15	1.41	20
Math	2.28	1.15	20	1.84	0.73	20	1.50	1.48	20
Writing	2.31	0.66	20	1.90	0.64	20	−0.47	1.34	20

Results II: Comparison between ML Algorithms

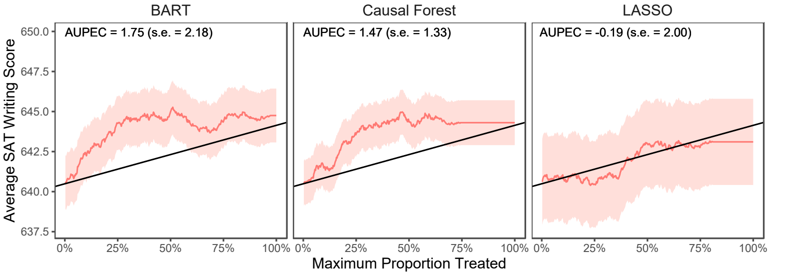
	Causal Forest				BART	
	vs. **BART**		vs. **LASSO**		vs. **LASSO**	
	est.	95% CI	est.	95% CI	est.	95% CI
Fixed ITR						
Math	1.55	[−0.35, 3.45]	1.83	[−0.50, 4.16]	0.28	[−2.39, 2.95]
Reading	1.86	[−0.79, 4.51]	1.31	[−1.49, 4.11]	−0.55	[−4.02, 2.92]
Writing	0.38	[−1.66, 2.42]	2.69	[−0.27, 5.65]	2.32	[−0.53, 5.15]
Estimated ITR						
Reading	−1.15	[−3.99, 1.69]	0.55	[−1.05, 2.15]	1.70	[−0.90, 4.30]
Math	−0.43	[−2.57, 3.43]	0.34	[−1.32, 2.00]	0.77	[−1.99, 3.53]
Writing	−0.41	[−1.63, 0.80]	2.37	[0.76, 3.98]	2.79	[1.32, 4.26]

Results III: AUPEC

Fixed ITR



Estimated ITR



Extension to Heterogeneous Treatment Effects

- Inference for heterogeneous treatment effects discovered via a generic ML algorithm
  - cannot assume ML algorithms converge uniformly
  - avoid computationally intensive method (e.g., repeated cross-fitting)
  - use Neyman's repeated sampling framework for inference
- Setup:
  - Conditional Average Treatment Effect (CATE):

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

  - CATE estimation based on ML algorithm

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

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

$$\tau_k \;:=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid c_{k-1}(s) \leq s(\mathrm{X}_i) < c_k(s))$$

    for $k = 1, 2, \ldots, K$ where c_k represents the cutoff between the $(k-1)$th and kth groups

GATE Estimation as ITR Evaluation

- A natural GATE estimator

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

where $\hat{f}_k(X_i) \;=\; 1\{s(X_i) \geq \hat{c}_k(s)\} - 1\{s(X_i) \geq \hat{c}_{k-1}(s)\}$

- Rewrite this as the PAPE:

$$\hat{\tau}_k \;=\; K\Bigg\{\underbrace{\frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i \hat{f}_k(X_i) + \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)(1-\hat{f}_k(X_i))}_{\text{estimated PAV}} \underbrace{- \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)}_{\text{no one gets treated}}\Bigg\}.$$

- Our results can be extended to both sample-splitting and cross-fitting

Concluding Remarks

- Use of ML algorithms is increasing in experimental studies
- Inference about ITRs has been largely model-based
  - We show how to experimentally evaluate ITRs
  - We incorporate budget constraints
  - No modeling assumption or asymptotic approximation is required
  - Complex ML algorithms can be used
  - Applicable to cross-fitting estimators
  - Simulations: good small sample performance
- Ongoing extensions
  - heterogeneous treatment effects using ML algorithms
  - dynamic ITRs
- Paper (JASA, forthcoming): https://arxiv.org/abs/1905.05389
- Software: evalITR available at CRAN