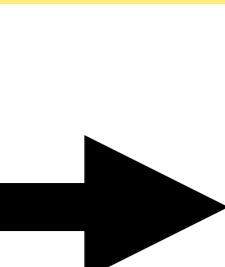


## Distributional Treatment Effect Modifiers

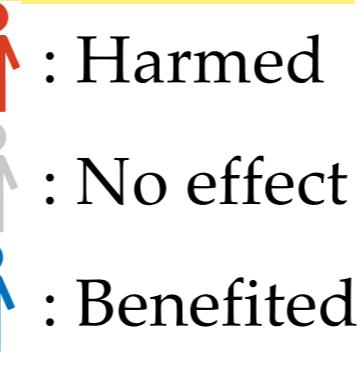
Yoichi Chikahara<sup>1,2</sup>, Makoto Yamada<sup>2</sup>, Hisashi Kashima<sup>2</sup><sup>1</sup> NTT

Motivation: Elucidate why treatment effects are different

## Treatment



## Outcome



- Harmed
- No effect
- Benefited

e.g., vaccination, education program

e.g., immunity, grades

Many existing methods use a complex ML model to accurately estimate heterogeneous treatment effects across individuals.

However, they offer no answer to the following question:

Different individuals have different treatment effects. Why?

We answer this question by solving the feature selection problem:

**Input:** Observations of features  $X$ , treatment  $A$ , and outcome  $Y$ **Output:** Features related to treatment effect heterogeneity

$X_1$	...	$X_d$	$A$	$Y$	$Y^1$	$Y^0$	$Y^1 - Y^0$
Male	15 y.o.	0	82	?	82	?	
Male	80 y.o.	0	174	?	174	?	
Female	64 y.o.	1	135	135	?	?	
Female	32 y.o.	1	110	110	?	?	

$p = 0.002$     $p = 0.019$     $p = 0.035$

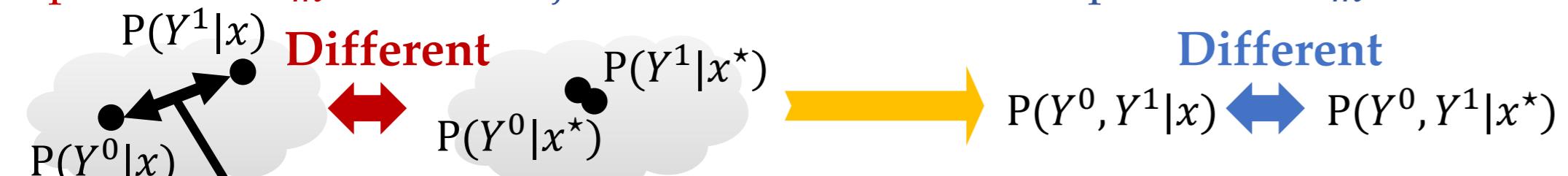
## Our Contributions

- Novel feature importance measure
- Its computationally efficient estimator
- Selection algorithm that controls Type I error

## Proposed method

**Our Goal:** Detect features whose values affect the functionals of joint distribution  $P(Y^0, Y^1 | X_m = x)$  (e.g., treatment effect variance)

## 1. Detecting distributional treatment effect modifiers

**Idea:** If the discrepancy between  $P(Y^0 | X_m = x)$  and  $P(Y^1 | X_m = x)$  depends on  $X_m = x$ , then joint distribution also depends on  $X_m = x$ .Measured by kernel MMD [2]:  $D_m^2(x) := \text{MMD}^2(P(Y^0 | X_m = x), P(Y^1 | X_m = x)) = \mathbb{E}_{Y^0, Y^0' | X_m = x} [k_Y(Y^0, Y^0')] + \mathbb{E}_{Y^1, Y^1' | X_m = x} [k_Y(Y^1, Y^1')] - 2 \mathbb{E}_{Y^0, Y^1 | X_m = x} [k_Y(Y^0, Y^1)]$ 

To detect the MMD value variation, we formulate feature importance measure as the variance of squared MMD:

$$I_m := \text{Var}[D_m^2(X_m)].$$

## 2. Estimating importance measure with IPW and RFFs

Using inverse probability weighting (IPW), we reformulate  $D_m^2(x)$  as

$$\begin{aligned} \text{WCMM}^2_{X_m=x} &:= \mathbb{E}_{A, A', X_m, X'_m, Y, Y' | X_m = x} [w^0(A, X) w^0(A', X') k_Y(Y, Y')] \\ &+ \mathbb{E}_{A, A', X_m, X'_m, Y, Y' | X_m = x} [w^1(A, X) w^1(A', X') k_Y(Y, Y')] \\ &- 2 \mathbb{E}_{A, A', X_m, X'_m, Y, Y' | X_m = x} [w^0(A, X) w^1(A', X') k_Y(Y, Y')]. \end{aligned}$$

$$\begin{aligned} w^0(A, X) &= \frac{\mathbf{I}(A = 0)}{1 - e(X)} \\ w^1(A, X) &= \frac{\mathbf{I}(A = 1)}{e(X)} \\ e(X) &:= P(A = 1 | X) \end{aligned}$$

Empirical estimator:  $\widehat{D}_m^2(x) := \sum_{i=1}^n \sum_{j=1}^n (\omega_i^{0,x} \omega_j^{0,x} + \omega_i^{1,x} \omega_j^{1,x}) k_Y(y_i, y_j) - 2 \sum_{i=1}^n \sum_{j=1}^n \omega_i^{0,x} \omega_j^{1,x} k_Y(y_i, y_j)$ If  $X_m$  is discrete,  $\omega_i^{a,x} = \frac{\mathbf{I}(x_{m,i} = x)}{\sum_{l=1}^n \mathbf{I}(x_{m,l} = x)} w^a(a_i, x_i)$ ; otherwise,  $\omega_i^{a,x} = \frac{1}{\sum_{l=1}^n k_{X_m}(x_{m,l}, x)} w^a(a_i, x_i)$ To reduce the computation time, we approximate  $k_Y$  with RFFs [3]:

$$k_Y(y_i, y_j) \approx \tilde{k}_Y(y_i, y_j) = \langle z(y_i), z(y_j) \rangle_{\mathbb{R}^r}, \quad z(y) = \begin{bmatrix} \sqrt{2} \cos(\lambda_1 y_1 + \zeta_1) \\ \vdots \\ \sqrt{2} \cos(\lambda_r y_r + \zeta_r) \end{bmatrix}$$

which yields

$$\widehat{D}_m^2(x) := \langle \tilde{\mu}_{Y^0|x}, \tilde{\mu}_{Y^0|x} \rangle_{\mathbb{R}^r} + \langle \tilde{\mu}_{Y^1|x}, \tilde{\mu}_{Y^1|x} \rangle_{\mathbb{R}^r} - 2 \langle \tilde{\mu}_{Y^0|x}, \tilde{\mu}_{Y^1|x} \rangle_{\mathbb{R}^r}.$$

Estimated feature importance:

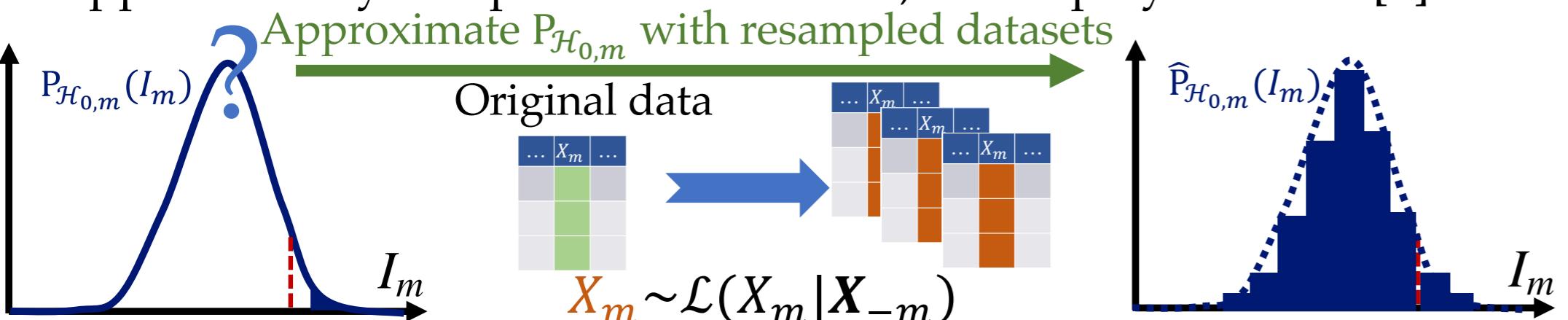
$$\widetilde{I}_m = \frac{1}{n-1} \sum_{i=1}^n \left( \widehat{D}_m^2(x_{m,i}) - \frac{1}{n} \sum_{s=1}^n \widehat{D}_m^2(x_{m,s}) \right)^2$$

## 3. Multiple tests with conditional randomization test (CRT)

We select features by performing multiple hypothesis tests:

$$\mathcal{H}_{0,m}: I_m = 0 \quad \text{and} \quad \mathcal{H}_{1,m}: I_m > 0. \quad (m=1, \dots, d)$$

To approximately compute the threshold, we employ the CRT [4]:



## Traditional mean-based approaches

Using the CATE conditioned on a single feature (i.e., the average treatment effect across individuals with identical attribute  $X_m = x$ ):

$$\begin{aligned} T_m(x) &:= \mathbb{E}[Y^1 - Y^0 | X_m = x] \\ &= \mathbb{E}[Y^1 | X_m = x] - \mathbb{E}[Y^0 | X_m = x], \end{aligned} \quad (1)$$

the existing methods (e.g., [1]) seek treatment effect modifiers:

**Definition 1** (Rothman et al. [2008]). Feature  $X_m$  is said to be a treatment effect modifier if there are at least two values of  $X_m$ ,  $x_m$  and  $x_m^*$  ( $x_m \neq x_m^*$ ), such that CATE  $T_m$  in (1) takes different values, i.e.,  $T_m(x_m) \neq T_m(x_m^*)$ .

## Weakness: Mean-based methods may overlook important features

## Example:

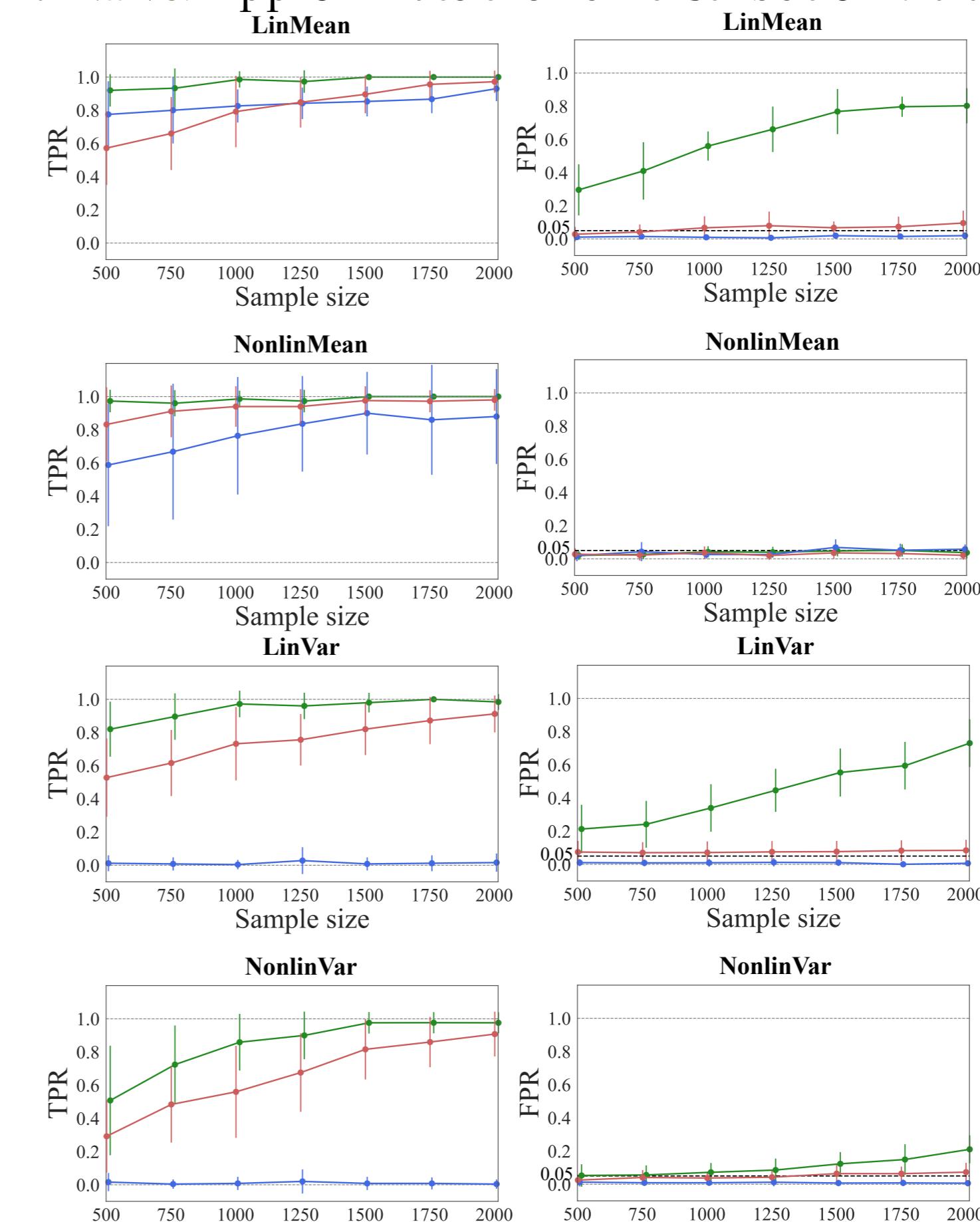
P( $Y^0, Y^1   X = 0$ )				P( $Y^0, Y^1   X = 1$ )					
$\backslash Y^1$	-1	0	1	Total	$\backslash Y^1$	-1	0	1	Total
-1	0	0	0	0	-1	0	0	0	0
0	0.5	0	0.5	1.0	0	0	1.0	0	1.0
1	0	0	0	0	1	0	0	0	0
Total	0.5	0	0.5	1.0	Total	0	1.0	0	1.0

Individuals with  $X=0$ Individuals with  $X=1$ 

$$\begin{aligned} Y^1 - Y^0: & \quad -1 \quad +1 \quad +1 \quad -1 \\ \mathbb{E}[Y^1 - Y^0 | X]: & \quad 0 \quad = \quad 0 \quad \text{Identical means} \\ \text{Var}[Y^1 - Y^0 | X]: & \quad 1 \quad \neq \quad 0 \quad \text{Different variances} \end{aligned}$$

## How can we detect distributional heterogeneity?

## Experimental results

**Synthetic data** We compare our method with the two baselines:1. **SI-EM** [1]: Mean-based approach2. **Naive**: Approximate the null distribution via a naive bootstrap**Naive** cannot control the FPR to  $\alpha = 0.05$ **SI-EM** cannot detect the features related to treatment effect variance

## Proposed achieves high TPR while controlling FPR

**Real-world data** We use health record dataset (from NHANES)Treatment  $A$ : obesity      Outcome  $Y$ : low-grade systemic inflammation  
Features  $X$ : e.g., age, gender, race, past medical history (e.g., asthma, stroke)

Feature	Adjusted p-value
Age	$0.0075 \pm 0.0305$
Gender	$0.0046 \pm 0.0269$
Number of cigarettes smoked	$0.0 \pm 0.0$

Not detected by SI-EM

[1] Qingyuan Zhao, Dylan S. Small, and Ashkan Ertefaie. "Selective inference for effect modification via the lasso". Journal of Royal Statistical Society: Series B (Statistical Methodology), 84(2):382–413, 2022.

[2] Arthur Gretton, Karsten M. Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. "A kernel two-sample test". JMLR, 13(1):723–773, 2012.

[3] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In NeurIPS, volume 3, page 5, 2007.

[4] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. "Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection". Journal of Royal Statistical Society: Series B (Statistical Methodology), 80 (3):551–577, 2018.