ML-assisted Randomization Tests for Causal Inference in Experimental Studies

Panos Toulis, Booth School of Business, University of Chicago

23rd Conference on Research on Economic Theory and Econometrics, Tinos, Greece, 2025

Collaborating Students

Wenxuan Guo (Booth)

JungHo Lee (CMU)

RAND Health Insurance Experiment

- The RAND Health Insurance Experiment (1971–1986) was a large-scale randomized study involving over 5,800 participants across six U.S. cities.
- Participants were randomly assigned healthcare plans: ranging from free to high copayments.
- **Result #1:** Cost sharing reduced healthcare spending without harming health outcomes (Brook et al., 2006).
- **Result #2:** Most studies also found treatment effect heterogeneity, particularly for minorities and low-income individuals (Newhouse, 1993; Aron-Dine et al., 2013).

Very influential for U.S. healthcare policy. But are these results reliable?

Heterogeneous treatment effects

Common regression approach:

$$\text{Spending}_i = \alpha + \beta \text{Treat}_i + \gamma \text{Treat}_i \times \text{Income}_i + ...$$

Estimating the average treatment effect is easy in a randomized study; e.g., regress $\text{Spending} \sim \text{Treat}$.

However, estimating heterogeneous effects through regression is tricky:

- $\text{Treat} \times \text{Income}$ is *not* randomized.
- OLS inference relies on asymptotics. But which regime is appropriate?
- The RAND experiment **is complex**. It has 4 time periods and plans are randomized on a family level.
- Do we "cluster" by individual, family, city, or time period?
- "Causal ML" methods are more flexible but also brittle..

Our Contribution

- Reliably estimating complex heterogeneous effects is challenging.
- We show that reliable results are possible if we lower our expectations, and aim to simply *test whether* certain types of heterogeneous effects exist.

  "If you want a marriage to last, look for someone with low expectations." — Warren Buffett

Key idea: use Machine Learning (ML) within standard Fisher Randomization Tests (FRT).

- Retains finite-sample validity of FRTs.
- Improved sensitivity through ML (as opposed to linear regression).
- Can test for heterogeneous effects, spillover effects, covariate balance, and more.

Outline

- Setup and Notation
- Main Procedure
- Type II error analysis
- Applications: global null, heterogeneity, spillovers (if time)
- Simulations

Setup

- $D = (D_1, \ldots, D_n) \in \{0,1\}^n$: binary *treatments*.
- Treatment assignment is known: $D \sim \mathbb{P}_n(D)$.
- $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$: *outcomes*.
- $X_1, \ldots, X_n$: *covariates*, $X_i \in \mathbb{R}^p$. Use $\mathbf{X} \in \mathbb{R}^{n \times p}$ for entire matrix.

We posit the outcome model:

$$Y_i = \mu + \underbrace{b(X_i)}_{\text{baseline}} + D_i \underbrace{h(X_i)}_{\text{direct effect}} + \underbrace{g(\mathbf{X}_{-i}, D_{-i})}_{\text{spillover}} + \varepsilon_i. \tag{1}$$

- b, h, g are arbitrary functions.
- ε_i is independent noise with $\mathbb{E}(\varepsilon_i \mid \mathbf{X}) = 0$, $\varepsilon \perp\!\!\!\perp D \mid \mathbf{X}$.

(1) follows Causal ML literature (Hill, 2011; Chernozhukov et al., 2018; Künzel et al., 2019).

1. Null hypothesis of no treatment effect

$$Y_i = \mu + \underbrace{b(X_i)}_{\text{baseline}} + D_i \underbrace{h(X_i)}_{\text{direct effect}} + \underbrace{g(\mathbf{X}_{-i}, D_{-i})}_{\text{spillover}} + \varepsilon_i.$$

As a starting point, consider the null:

$$H_0^{\text{glob}} : h = 0, g = 0 \quad \text{v.s.} \quad H_1^{\text{glob}} : h \neq 0, g = 0.$$

- Under the potential outcomes framework, $H_0^{\text{glob}} \equiv Y_i(1) \stackrel{d}{=} Y_i(0)$. This is slightly weaker than "Fisher's sharp null".
- i.e., H_0^{glob} implies no treatment effect whatsoever.

ML-based test statistic

To test H_0^{glob}, we propose constructing two models using ML:

$$\mathcal{M}_0 : Y_i \sim f(X_i),$$
$$\mathcal{M}_1 : Y_i \sim f(D_i, X_i).$$

f = neural nets, trees, penalized regression, etc.

Define the test statistic as

$$t_n(Y, D, \mathbf{X}) := \mathrm{CV}_{n,k}(\mathcal{M}_0) - \mathrm{CV}_{n,k}(\mathcal{M}_1), \tag{2}$$

where $\mathrm{CV}_{n,k}(\mathcal{M})$: k-fold cross-validated squared loss of model $\mathcal{M}$.

- Intuitively, $t_n(Y, D, \mathbf{X})$ measures whether D is predictive of Y.
- This test statistic allows for complex dependencies between treatment and outcome, as captured by $\mathcal{M}_1$.

An ANOVA-like idea

- Equation (2) describes an ANOVA-type statistic.
- e.g., Gerber and Green (2012) used a similar statistic based on classical ANOVA:

$$\mathrm{SSR}(\mathcal{M}_0) - \mathrm{SSR}(\mathcal{M}_1) \ .$$

  Difference: our method uses ML instead of linear models.

- We also point out similarities to **variable importance** measures in statistics and machine learning (Breiman, 2001; Strobl et al., 2008; Williamson et al., 2021; Bénard et al., 2022).

$$\text{“importance of } D\text{”} = \mathbb{E}[Y - \mathbb{E}(Y \mid X)]^2 - \mathbb{E}[Y - \mathbb{E}(Y \mid X, D)]^2 \ .$$

- Our test's power can be characterized by a similar measure.

Main testing procedure

Procedure 1 (ML-assisted Randomization Test)

1. Obtain observed value of $T_n = t_n(Y, D, \mathbf{X})$ as defined in (2).
2. Compute $t^{(r)} = t_n(Y, D^{(r)}, \mathbf{X})$, $D^{(r)} \overset{iid}{\sim} \mathbb{P}_n$, for $r = 1, \ldots, R$.
3. Calculate p-value:

$$\text{pval} = \frac{1}{1+R}\left[\sum_{r=1}^{R} \mathbb{1}\{t^{(r)} > T_n\} + 1\right]. \tag{3}$$

- The test is finite-sample valid (Lehmann and Romano, 2005):

$$\mathbb{P}\big(\text{pval} \leq \alpha\big) \leq \alpha, \text{ for any } \alpha \in [0,1] \text{ and any } n > 0.$$

Proof:

$$\underbrace{t(Y, D^{(r)}, \mathbf{X})}_{\text{randomization distr.}} \overset{H_0}{=} t(Y^{(r)}, D^{(r)}, \mathbf{X}) \overset{d}{=} t(Y, D, \mathbf{X}) := \underbrace{T_n}_{\text{sampling distr.}}.$$

Inference through test inversion

- Caveat: Procedure 1 only detects a treatment effect, but it does not quantify it ("omnibus test").
- Still, inference is possible through **test inversion** under assumptions.
- In particular, over a grid of values for μ, we can use Procedure 1 to test

$$H_0 : Y_i(1) \stackrel{d}{=} \mu + Y_i(0).$$

- Then, $\mathrm{CI} = \{\text{non-rejected } \mu\}$ is the randomization-based confidence interval.
- This interval is also **valid in finite samples**, assuming a constant treatment effect.

Type II error bound

Theorem

Suppose certain assumptions and additional regularity conditions on $\mathcal{F}, \mathcal{F}_0$ hold with $k = O(1)$. Then, under the alternative H_1^{glob}, for some small constant $C > 0$,

$$\mathbb{P}(\text{pval} > \alpha) = O\left(k \exp\left(-\frac{Cn\Delta^2}{kM^4}\right)\right),$$

- $\mathcal{F}_0, \mathcal{F}$ are function classes for $\mathcal{M}_0, \mathcal{M}_1$, respectively.
- Quantity Δ measures the *variable importance* of treatment:

$$\Delta := \underbrace{\inf_{f_0 \in \mathcal{F}_0} \mathbb{E}(Y - f_0(X))^2}_{\text{prediction error in the reduced model}} - \underbrace{\inf_{f \in \mathcal{F}} \mathbb{E}(Y - f(X, D))^2}_{\text{prediction error in the full model}}.$$

- e.g., $\Delta = \pi(1-\pi)\tau^2$, in a linear model $y = a + bx + \tau d$.

Takeaway: better prediction $\Rightarrow$ larger Δ $\Rightarrow$ higher power!

2. Null hypothesis on treatment effect heterogeneity

Recall:

$$Y_i = \mu + \underbrace{b(X_i)}_{\text{baseline}} + D_i \underbrace{h(X_i)}_{\text{direct effect}} + \underbrace{g(\mathbf{X}_{-i}, D_{-i})}_{\text{spillover}} + \varepsilon_i.$$

Under this outcome model, we may define the null hypothesis of homogeneous treatment effects as

$$H_0^{\text{het}} : h(x) = \tau, g = 0 \quad \text{v.s.} \quad H_1^{\text{het}} : h(x) \neq \tau, g = 0,$$

for some unknown constant $\tau \in \mathbb{R}$. The null implies that

$$Y_i = \mu + b(X_i) + \tau D_i + \varepsilon_i.$$

Procedure 1 **is not valid** because we now allow for a (first-order) treatment effect.

Main idea: majorize the test over a grid for τ.

Procedure for testing heterogeneity

1. Transform $Y_i^0 = Y_i - \tau_0 D_i$ for some $\tau_0 \in \mathbb{R}$.
2. Apply Procedure 1 on $\{(Y_i^0, D_i, X_i)\}_{i=1}^n$ to obtain $\mathrm{pval}(\tau_0)$.
3. Maximize the p-values over a grid of values for τ_0:

$$\mathrm{pval}_{\mathrm{het}} = \sup_{\tau_0 \in \mathbb{R}} \mathrm{pval}(\tau_0). \tag{4}$$

- That is, we test whether the residual $Y_i - \tau D_i$ is "fully explained" by covariates $\mathbf{X}$ for some fixed τ. Bread-and-butter for ML.
- The test based on (4) is also finite-sample valid due to majorization over valid p-values (Berger and Boos, 1994).

Application: RAND Health Insurance Experiment

- We begin by testing the global null hypothesis H_0^{glob} using Procedure 1 (no treatment effect).
- The randomization test **strongly rejects** this hypothesis ($p \approx 0$).
- By inverting the test, we also obtain **[0.51, 0.72]** as the 95% randomization-based confidence interval.
- Roughly speaking, this suggests that cost-sharing reduces medical expenses by **54.1%**, *assuming* the effect is constant.
- These results are in agreement with earlier work.

Application: RAND Health Insurance Experiment

- However, we find **no treatment effect heterogeneity.**
- Majorized p-values are shown below:

	Synthetic data[1]	Real data
Simple linear model	0.428	0.214
Linear model + interactions	0.040	0.752
Tree models (via `xgboost`)	0.000	0.424

- In the real data, all ML-assisted tests do not reject that $Y - \tau D$ is fully explained by $\mathbf{X} \Rightarrow$ No evidence for treatment heterogeneity.
- This contrasts with most prior analyses that have overwhelmingly suggested heterogeneity.
- Conjecture: prior methods relied on asymptotic OLS inference and perhaps mistakenly attributed heterogeneity to spurious correlations between D and X (which exist in the RAND data.)

[1] Created based on original RAND study, but with synthetic outcomes designed to produce heterogeneous effects.

Limitations and future directions

- Our tests do not quantify the treatment in the traditional econometric sense (only in a predictive sense).
- They are best suited as *screening / diagnostic tools*. Possibly synergistic with classical OLS.
- Future work: scalability by online learning on the test statistic.
- Include complex designs such as factorial experiments.

Thank you!

- Wenxuan Guo, JungHo Lee, and Panos Toulis, "ML-assisted Randomization Tests for Detecting Treatment Effects in A/B Experiments ," `https://arxiv.org/abs/2501.07722`, 2025.

References

Athey, S., Eckles, D., and Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240.

Basse, G., Ding, P., Feller, A., and Toulis, P. (2024). Randomization tests for peer effects in group formation experiments. *Econometrica*, 92(2):567–590.

Basse, G., Feller, A., and Toulis, P. (2019). Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494.

Bénard, C., Da Veiga, S., and Scornet, E. (2022). Mean decrease accuracy for random forests: inconsistency, and a practical solution via the sobol-mda. *Biometrika*, 109(4):881–900.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Gerber, A. and Green, D. (2012). *Field Experiments: Design, Analysis, and Interpretation*. W. W. Norton.

Hennessy, J., Dasgupta, T., Miratrix, L., Pattanayak, C., and Sarkar, P. (2016). A conditional randomization test to account for covariate imbalance in randomized experiments. *Journal of Causal Inference*, 4(1):61–80.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, 9:1–11.

Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22.

3. Testing for spillover effects (short)

We test

$$H_0^{\text{sp}} : g = 0, \quad \text{v.s.} \quad H_1^{\text{sp}} : g \neq 0.$$

We use the models:

$$\mathcal{M}_0^{\text{sp}} : Y_i \sim f(D_i, X_i),$$
$$\mathcal{M}_1^{\text{sp}} : Y_i \sim f(D_i, X_i, A_{i\cdot}^\top D).$$

- $\mathbf{A} \in \{0,1\}^{n \times n}$: observed adjacency matrix between units.
- $A_{i\cdot}^\top D =$ "number of treated neighbors" of i.
- If spillover effects exist (i.e., $g \neq 0$), these will be partially captured by $\mathcal{M}_1$ but not by $\mathcal{M}_0$.
- We need to adapt Procedure 1 in one more crucial way to account for interference through the use of **conditional** FRTs (Athey et al., 2018; Basse et al., 2019, 2024). See paper.

Test for covariate balance

- *Covariate balance* refers to how similar the covariate distributions are between the treated and control groups.
- In balanced experiments, any observed differences between treated and control can be attributed to the treatment effect.

We can test for covariate balance by modifying the fitted models:

$$\mathcal{M}_0^{\text{im}} : X^j \sim \mathbf{X}^{-j},$$
$$\mathcal{M}_1^{\text{im}} : X^j \sim Z + \mathbf{X}^{-j},$$

where X^j denotes the j-th covariate and $\mathbf{X}^{-j}$ the sub-matrix of $\mathbf{X}$ without the j-th covariate.

A *significant p-value* indicates covariate imbalance in X^j.

Adjust for covariate imbalance

In case the above test rejects, Procedure 1 may be adjusted in a way that can increase the power of the randomization test.

- $\mathrm{pval}_j(z, \mathbf{X})$: p-value from the covariate balance ML-FRT.
- $q := \mathrm{pval}_j(Z, \mathbf{X})$, the realized value in the experiment.

We then modify Procedure 1 by replacing Step 2 with:

2'. Compute the randomized statistic $t^{(r)} = t_n(Y, Z^{(r)}, \mathbf{X})$, $Z^{(r)} \overset{iid}{\sim} \mathbb{P}^j_n$, for $r = 1, \ldots, R$.

$\mathbb{P}^j_n(z) \propto \mathbb{1}\{\mathrm{pval}_j(z, \mathbf{X}) \leq q\} \times \mathbb{P}_n(z)$ is the conditional randomization distribution, given that the imbalance p-value less or equal to q.

The conditional randomization test under Step 2' remains finite-sample valid with improved power Hennessy et al. (2016, Section 5.2).