Research Article

Michael Lingzhi Li* and Kosuke Imai

Neyman meets causal machine learning: Experimental evaluation of individualized treatment rules

https://doi.org/10.1515/jci-2023-0072
received October 21, 2023; accepted April 22, 2024

Abstract: A century ago, Neyman showed how to evaluate the efficacy of treatment using a randomized experiment under a minimal set of assumptions. This classical repeated sampling framework serves as a basis of routine experimental analyses conducted by today's scientists across disciplines. In this article, we demonstrate that Neyman's methodology can also be used to experimentally evaluate the efficacy of individualized treatment rules (ITRs), which are derived by modern causal machine learning (ML) algorithms. In particular, we show how to account for additional uncertainty resulting from a training process based on cross-fitting. The primary advantage of Neyman's approach is that it can be applied to any ITR regardless of the properties of ML algorithms that are used to derive the ITR. We also show, somewhat surprisingly, that for certain metrics, it is more efficient to conduct this *ex-post* experimental evaluation of an ITR than to conduct an *ex-ante* experimental evaluation that randomly assigns some units to the ITR. Our analysis demonstrates that Neyman's repeated sampling framework is as relevant for causal inference today as it has been since its inception.

Keywords: causal inference, machine learning, individualized treatment rule, policy evaluation, repeated sampling

MSC 2020: 62G05

1 Introduction

Neyman's seminal 1923 paper introduced two foundational ideas in causal inference [1]. First, Neyman developed a formal notation for potential outcomes and defined the average treatment effect (ATE) as a causal quantity of interest. Second, he showed how randomization of treatment assignment alone can be used to establish the unbiasedness and estimation uncertainty of the standard difference-in-means estimator. Since then, combined with the additional assumption of random sampling of units, Neyman's repeated sampling framework has served as a basis of routine experimental analyses conducted by scientists across many disciplines.

Over the past two decades, however, the causal inference literature has gone beyond the ATE. Specifically, the realization that the same treatment can have varying impacts on different individuals led to the development of statistical methods and machine learning (ML) algorithms for estimating heterogeneous treatment effects (e.g., [2–5]). Furthermore, a number of researchers have developed various methods for deriving data-driven

*** Corresponding author: Michael Lingzhi Li,** Technology and Operations Management Unit, Harvard Business School, Boston MA, 02163, USA, e-mail: mili@hbs.edu
Kosuke Imai: Department of Statistics and Department of Government, Harvard University, Cambridge MA, 02138, USA, e-mail: imai@harvard.edu
ORCID: Michael Lingzhi Li 0000-0002-2456-4834; Kosuke Imai 0000-0002-2748-1022

individualized treatment rules (ITRs) (e.g., [6–13]). With an increasing availability of granular data and modern computing power, these ITRs are becoming popular in business, medicine, politics, and even public policy.

In this article, we demonstrate that Neyman's repeated sampling framework is still relevant for today's causal ML methods. We show how the framework can be used to experimentally evaluate the efficacy of any ITRs (including those obtained with ML algorithms via cross-fitting) under a minimal set of assumptions. While some of our formal results are originally derived in our previously published work [14] or follow directly from them, we focus on the intuition behind those theoretical results to facilitate the future extensions to other settings.

We also show, using Neyman's framework, that it is not always statistically more efficient to evaluate an ITR by conducting a new randomized experiment where the treatment is the administration of the ITR itself (i.e., *ex-ante* evaluation) than simply using the data from an existing randomized controlled trial (i.e., *ex-post* evaluation). Altogether, this article shows how Neyman's classical methodological framework can be applied to solve today's causal inference problems.

2 Neyman's repeated sampling framework

We begin by briefly introducing Neyman's inferential approach to estimating the ATE. Suppose that we have a sample of n units, and for each unit, we define two potential outcomes, $Y_i(1)$ and $Y_i(0)$, under the treatment and control conditions, respectively. Let T_i denote the binary treatment assignment variable, which is the ith element of n-dimensional treatment vector $\boldsymbol{T}$. Then, the observed outcome can be written as $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$. Finally, $\boldsymbol{X}_i$ denotes a set of observed pre-treatment covariates for unit i, where $\mathcal{X}$ represents the support of covariate distribution.

As pointed out by Rubin in his discussion of Neyman's 1923 paper [15], the aforementioned setup implicitly assume no interference between units – the outcome of one unit is not influenced by the treatment of another unit. We explicitly state this assumption below.

Assumption 1. (No interference between units) The potential outcomes for unit i do not depend on the treatment status of other units, i.e., for all $t_1, t_2, \ldots, t_n \in \{0, 1\}$, we have

$$Y_i(T_1 = t_1, T_2 = t_2, \ldots, T_n = t_n) = Y_i(T_i = t_i).$$

Neyman considered the classical randomized experiment where the treatment assignment is completely randomized with n_1 units assigned to the treatment condition and the remaining $n_0 = n - n_1$ units assigned to the control condition.

Assumption 2. (Complete randomization of treatment assignment) The treatment assignment probability is given by

$$\Pr(\boldsymbol{T} = \boldsymbol{t} \mid \{Y_i(1), Y_i(0), \boldsymbol{X}_i\}_{i=1}^n) = \frac{1}{\binom{n}{n_1}},$$

for each $\boldsymbol{t}$, where $\sum_{i=1}^n t_i = n_1$.

Under these two assumptions alone, Neyman showed the following sample average treatment effect (SATE) can be estimated without bias:

$$\tau_{\text{SATE}} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}.$$

Using the difference-in-means estimator $\hat{\tau}$,

$$\mathbb{E}(\hat{\tau} \mid \{Y_i(1), Y_i(0)\}_{i=1}^n) = \tau_{\text{SATE}}, \quad \text{where } \hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} T_iY_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_i)Y_i.$$

Neyman also showed that the variance of this estimator is not identifiable but a conservative variance can be estimated from the data without bias:

$$\mathbb{V}(\hat{\tau} \mid \{Y_i(1), Y_i(0)\}_{i=1}^n) = \frac{1}{n}\left(\frac{n_0}{n_1}S_1^2 + \frac{n_1}{n_0}S_0^2 + 2S_{01}\right) \leq \frac{1}{n}\left(\frac{n_0}{n_1}S_1^2 + \frac{n_1}{n_0}S_0^2 + (S_1^2 + S_0^2)\right) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0},$$

where $S_t^2 = \sum_{i=1}^n (Y_i(t) - \overline{Y(t)})^2/(n-1)$, $S_{01} = \sum_{i=1}^n (Y_i(0) - \overline{Y(0)})(Y_i(1) - \overline{Y(1)})/(n-1)$, and $\overline{Y(t)} = \sum_{i=1}^n Y_i(t)/n_t$ for $t = 0, 1$.

Neyman obtained the aforementioned results by averaging over all possible treatment assignments under complete randomization. Subsequent work has extended Neyman's framework to a superpopulation framework by assuming that the sample of n units are obtained, through random sampling, from a superpopulation of infinite size $\mathcal{P}$.

Assumption 3. (Random sampling of units) Each of n units, represented by a three-tuple consisting of two potential outcomes and pre-treatment covariates, is assumed to be independently sampled from a superpopulation $\mathcal{P}$, i.e.,

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

This extended framework, which we call Neyman's repeated sampling framework, is useful because it allows us to estimate the population ATE from the sample,

$$\tau_{\text{PATE}} = \mathbb{E}(Y_i(1) - Y_i(0)).$$

Subsequent work has shown that the difference-in-means estimator is unbiased for the PATE and the exact variance can be estimated without bias [16]:

$$\mathbb{E}(\hat{\tau}) = \mathbb{E}[\mathbb{E}(\hat{\tau} \mid \{Y_i(1), Y_i(0)\}_{i=1}^n)] = \mathbb{E}[\tau_{\text{SATE}}] = \tau_{\text{PATE}}, \tag{1}$$

$$\mathbb{V}(\hat{\tau}) = \mathbb{E}[\mathbb{V}(\hat{\tau} \mid \{Y_i(1), Y_i(0)\}_{i=1}^n)] + \mathbb{V}[\mathbb{E}(\hat{\tau} \mid \{Y_i(1), Y_i(0)\}_{i=1}^n)] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}, \tag{2}$$

where $\sigma_t^2 = \mathbb{V}(Y_i(t))$ for $t = 0, 1$.

In the remainder of this article, we will show that this Neyman's repeated sampling framework enables an assumption-free experimental evaluation of data-driven ITRs.

3 Experimental evaluation of ITRs

In this section, we explain how Neyman's repeated sampling framework can be applied to experimentally evaluate the empirical performance of ITRs, which assigns each individual unit to either the treatment or control condition based on their observed characteristics.

3.1 Setup

Suppose that we use an ML algorithm to create a ITR,

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

Most commonly, researchers first estimate the conditional ATE [see, e.g., 7,17–21]:

$$\tau(\boldsymbol{x}) \coloneqq \mathbb{E}[Y_i(1) - Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}].$$

They then derive an ITR as the treatment rule that assigns the treatment to everyone who is predicted to have a positive CATE, i.e., $f(\boldsymbol{x}) = \mathbf{1}\{\hat{\tau}(\boldsymbol{x}) > 0\}$, where $\hat{\tau}(\boldsymbol{x})$ is an estimate of the CATE. Throughout this article, we assume, without loss of generality, that a positive effect implies that the treatment is beneficial. One may also

consider a cost associated with the administration of treatment or a budget constraint that limits the proportion of individuals who can receive the treatment.

Our goal is to evaluate the empirical performance of an ITR without assuming that an ITR is indeed optimal. In the aforementioned example, we do not assume that $\hat{\tau}(\boldsymbol{x})$ is an accurate estimate of the CATE. In fact, we do not make any assumption about how the ITR is constructed and how accurate it is. The ITR may be derived from an application of an ML algorithm or be even based on heuristics. For now, we only assume that the ITR to be evaluated is given. For example, it may be estimated from an external dataset to be used in a downstream decision-making context. In Section 5, we discuss how to use the same experimental data for both learning and evaluating an ITR.

To measure the performance of an ITR, we consider two quantities. The first is the population average value (PAV), which is defined as

$$\lambda_f \coloneqq \mathbb{E}[Y_i(f(\boldsymbol{X}_i))]. \tag{3}$$

This is the standard metric of ITR's overall performance. The second quantity is the population average prescriptive effect (PAPE) [14,22], which measures the benefit of ITR and is defined as follows:

$$\tau_f \coloneqq \mathbb{E}[Y_i(f(\boldsymbol{X}_i))] - p_f\mathbb{E}[Y_i(1)] - (1 - p_f)\mathbb{E}[Y_i(0)], \tag{4}$$

where $p_f = \Pr(f(\boldsymbol{X}_i) = 1)$ represents the proportion of individuals who are treated by the ITR f.

The PAPE compares the performance of ITR against the non-ITR that treats the same proportion of randomly selected individuals. This contrasts with other quantities considered in the literature such as the targeting operator characteristic (TOC) [23], which compares the performance of ITR against the non-individualized rule that treats everyone. Unlike the TOC, the PAPE focuses on the benefit of determining which individuals should be treated while holding the proportion of those who receive the treatment constant.

To gain additional intuition about the PAPE, consider the following alternative but equivalent expression of the same quantity:

$$\tau_f = \mathrm{Cov}(f(\boldsymbol{X}_i), Y_i(1) - Y_i(0)).$$

This alternative expression shows that the PAPE measures how well the ITR agrees with the true individual treatment effect (ITE). To compare across datasets, we can further normalize the PAPE as the correlation between the ITR and the true ITE, i.e.,

$$\frac{\tau_f}{\sqrt{\mathbb{V}(f(\boldsymbol{X}_i))\mathbb{V}(Y_i(1) - Y_i(0))}} = \mathrm{Corr}(f(\boldsymbol{X}_i), Y_i(1) - Y_i(0)).$$

Although this provides a scale-invariant quantity to understand the performance of ITR, it is not identifiable from the data because we cannot identify the variance of ITE, i.e., $\mathbb{V}(Y_i(1) - Y_i(0))$.

The aforementioned equality further implies the following inequality by applying Cauchy-Schwarz twice:

$$\tau_f \le \sqrt{\mathbb{V}(f(\boldsymbol{X}_i))\mathbb{V}(Y_i(1) - Y_i(0))} \le \sqrt{2p_f(1 - p_f)(\mathbb{V}(Y_i(1)) + \mathbb{V}(Y_i(0)))}. \tag{5}$$

Therefore, PAPE is bounded provided that the second moments of the potential outcomes exist. Thus, given a fixed variance of the potential outcomes, τ_f is most likely largest around $p_f = 0.5$. This is because when p_f is around 0.5, the ITR has the greatest room to deviate from the randomized treatment rule.

3.2 Estimation and inference

To estimate the PAV and PAPE under Neyman's repeated sampling framework, we consider the following "difference-in-means"-type estimators:

$$\hat{\lambda}_f(\boldsymbol{Z}_n) = \frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i f(\boldsymbol{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1 - T_i)(1 - f(\boldsymbol{X}_i)), \tag{6}$$

$$\hat{\tau}_f(\boldsymbol{Z}_n) = \frac{n}{n-1}\left[\frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i f(\boldsymbol{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)(1-f(\boldsymbol{X}_i)) - \frac{\hat{p}_f}{n_1}\sum_{i=1}^{n} Y_i T_i - \frac{1-\hat{p}_f}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)\right], \tag{7}$$

where $\hat{p}_f = \sum_{i=1}^{n} f(\boldsymbol{X}_i)/n$ is the estimated population proportion of individuals who receive the treatment assignment and $\boldsymbol{Z}_n = \{Y_i, T_i, \boldsymbol{X}_i\}_{i=1}^{n}$ represents the experimental data of sample size n. The factor $n/(n-1)$ in the PAPE estimator represents the loss in one degree-of-freedom due to estimating the population-level quantity p_f.

The following theorems, reproduced from our previously published work [14], show that, under Neyman's repeated sampling framework, these two estimators are unbiased and the finite-sample variances can be derived.

Theorem 1. (Unbiasedness and variance of the PAV estimator [14]) Under *Assumptions* 1–3, *the expectation and variance of the PAV estimator defined equation* (6) *are given by*

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

where $S_{ft}^2 = \sum_{i=1}^{n}(Y_{fi}(t) - \overline{Y_f(t)})^2/(n-1)$ *with* $Y_{fi}(t) = \mathbf{1}\{f(\boldsymbol{X}_i) = t\}Y_i(t)$, *and* $\overline{Y_f(t)} = \sum_{i=1}^{n} Y_{fi}(t)/n$, *for* $t = \{0, 1\}$.

Theorem 2. (Unbiasedness and variance of the PAPE estimator [14]) *Under Assumptions* 1–3, *the expectation and variance of the PAPE estimator defined equation* (7) *are given by*,

$$\mathbb{E}\{\hat{\tau}_f(\boldsymbol{Z}_n)\} = \tau_f,$$
$$\mathbb{V}\{\hat{\tau}_f(\boldsymbol{Z}_n)\} = \frac{n^2}{(n-1)^2}\left[\frac{\mathbb{E}(\widetilde{S}_{f1}^2)}{n_1} + \frac{\mathbb{E}(\widetilde{S}_{f0}^2)}{n_0} + \frac{1}{n^2}\{\tau_f^2 - np_f(1-p_f)\tau^2 + 2(n-1)(2p_f-1)\tau_f\tau\}\right],$$

where $\widetilde{S}_{ft}^2 = \sum_{i=1}^{n}(\widetilde{Y}_{fi}(t) - \overline{\widetilde{Y}_f(t)})^2/(n-1)$ *with* $\widetilde{Y}_{fi}(t) = (f(\boldsymbol{X}_i) - \hat{p}_f)Y_i(t)$, *and* $\overline{\widetilde{Y}_f(t)} = \sum_{i=1}^{n}\widetilde{Y}_{fi}(t)/n$, *for* $t = \{0, 1\}$.

The properties of the PAV estimator shown above follow immediately from Neyman's classic results by replacing the potential outcome $Y_i(t)$ with the the potential outcome that incorporates the ITR, i.e., $\mathbf{1}\{f(\boldsymbol{X}_i) = t\}Y_i(t)$. The form of the estimator does not mean, however, that we are ignoring observations whose prescribed treatment status differs from the observed one, i.e., $f(\boldsymbol{X}_i) \neq T_i$. These observations are still used when estimating the variance. An important implication of this subtle fact is discussed in Section 3.4.

We can also compare the results of the PAPE estimator with Neyman's classic results. We observe that the relevant potential outcome is given by $\widetilde{Y}_{fi}(t) = (f(\boldsymbol{X}_i) - \hat{p}_f)Y_i(t)$, which directly compares the ITR with the randomized treatment rule. When compared to the PAV estimator, the variance of the PAPE estimator has an additional term, which comes from the fact that $\widetilde{Y}_{fi}(t)$ is correlated across observations due to the estimation of the proportion p_f.

The correlation can be broadly decomposed into two components: (1) a negative component caused by the negative correlation within the mean-adjusted ITR ($f(\boldsymbol{X}_i) - \hat{p}_f$) and ($f(\boldsymbol{X}_j) - \hat{p}_f$), and (2) the remaining component that arises due to the interaction between $f(\boldsymbol{X}_i)$ and τ_i. The negative component generally dominates if the ITR of interest treats roughly 50% of the population ($p_f \approx 0.5$) and the ATE τ is not too small. This suggests that the mean adjustment could lead to a variance reduction when the ITR treats a roughly half of the population.

On the other hand, for $p_f \neq 0.5$ and $n \gg 1$, this additional term is positive if and only if the following condition holds:

$$|\tau_f| \geq \frac{np_f(1-p_f)}{(n-1)|2p_f-1|}|\tau|.$$

Thus, this additional term is only likely to be positive under a scenario where p_f is away from 0.5 and the magnitude of the PAPE is much greater than that of the ATE (i.e., the ITR is performing well). Equation (5) suggests that this is unlikely unless the variance of the individualized treatment effect is large, implying a large degree of treatment effect heterogeneity.

3.3 Performance comparison among multiple ITRs

While the PAPE compares an ITR with a random treatment assignment rule that treats the same proportion of units, researchers are often interested in comparing the performance of multiple ITRs. In such cases, we recommend estimating the difference in PAV between two ITRs that are subject to the same budget constraint. Imai and Li [14] provided the details of estimation and inference regarding this quantity.

The use of PAPE for comparison of two ITRs is sometimes inappropriate. To see this, note that the difference in PAPE can be written as the covariance between the agreement of two ITRs and the true ITE:

$$\tau_f - \tau_g = \mathrm{Cov}(f(\boldsymbol{X}_i) - g(\boldsymbol{X}_i), Y_i(1) - Y_i(0)).$$

This expression shows that for ITRs with similar treatment proportions, the sign of this difference indicates the relative capability of the ITRs in identifying the optimal individuals to treat. However, if the two ITRs f and g have significantly different treatment proportions, then this comparison is difficult. Figure 1 shows an example, in which ITR f has a higher PAV than ITR g, but f has a negative PAPE, while g has a positive PAPE. In this case, f is not an effective ITR as it performs significantly worse than random treatment, but practitioners might still choose to not use g as it is only able to identify a small percentage of good patients to target.

3.4 Lack of invariance

Unlike Neyman's ATE estimator, the PAV and PAPE estimators are not invariant to a constant shift of the outcome variable. One might expect that adding a constant δ to Y would shift the PAV estimator by δ and not affect the PAPE estimator at all. Unfortunately, this is not the case. For both of these estimators, a constant shift of the outcome will result in an *additional* change of the equal magnitude. Let $\hat{\lambda}_f^{\delta}(\boldsymbol{Z})$ and $\hat{\tau}_f^{\delta}(\boldsymbol{Z})$ be the new PAV and PAPE estimators under a constant δ shift, i.e., $\hat{\lambda}_f^{\delta}(\boldsymbol{Z}) = \hat{\lambda}_f(\boldsymbol{Z}) + \delta$ and $\hat{\tau}_f^{\delta}(\boldsymbol{Z}) = \hat{\tau}_f(\boldsymbol{Z}) + \delta$. Then, we have,

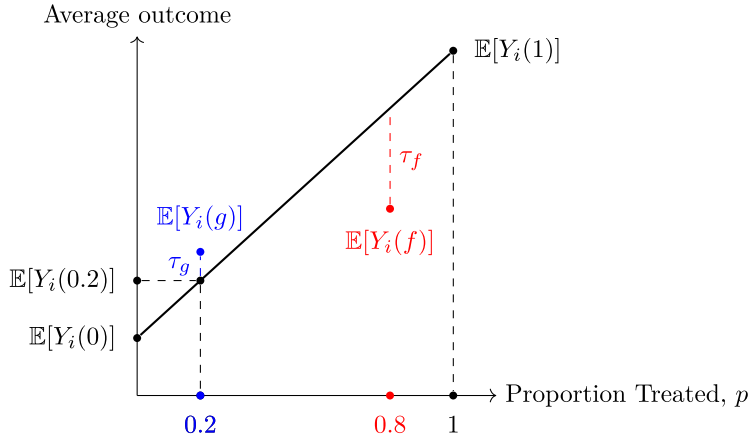

Figure 1: Illustration of PAPE for two different ITRs f and g. Here, the x axis is the proportion of individuals treated, and y axis is PAV. The PAV of f is higher than PAV of g, but ITR g has a positive PAPE τ_g and the ITR f has a negative PAPE τ_f.

$$\hat{\lambda}_f^{\delta}(\boldsymbol{Z}) - \hat{\lambda}_f(\boldsymbol{Z}) - \delta = \frac{n-1}{n}(\hat{\tau}_f^{\delta}(\boldsymbol{Z}) - \hat{\tau}_f(\boldsymbol{Z})) = \delta\left(\frac{1}{n_1}\sum_{i=1}^{n} T_i f(\boldsymbol{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)(1-f(\boldsymbol{X}_i)) - 1\right).$$

Since this term equals zero in expectation, the two estimators remain unbiased. However, this shift affects their variances because the ITR $f(\boldsymbol{X}_i)$ is only balanced on average due to the randomized treatment assignment. Intuitively, this is because the relevant potential outcomes in the case of both PAV and PAPE are not invariant to a constant shift of the outcome variable. For PAV, the relevant potential outcome is $Y_{fi}(t) = \mathbf{1}\{f(\boldsymbol{X}_i) = t\}Y_i(t)$, so under a constant shift of δ, only samples with $f(\boldsymbol{X}_i) = t$ would be shifted by δ, while the remaining samples still have a value of zero. The intuition is similar for PAPE. Therefore, we expect that balancing the observed outcomes Y_i close to zero could result in increased efficiency.

We now derive the constant shift δ that minimizes the resulting variances. We can show that a constant shift of δ to potential outcomes creates the following additional variance terms for both estimators:

Proposition 1. (Minimum variance estimators) *The variances of the constant shift estimators are given by*

$$\mathbb{V}(\hat{\lambda}_f^{\delta}(\boldsymbol{Z})) = \mathbb{V}(\hat{\lambda}_f(\boldsymbol{Z})) + \delta p_f(1-p_f)\left(\frac{2\kappa_{11}}{n_1} + \frac{2\kappa_{00}}{n_0} + \delta\cdot\frac{n}{n_1 n_0}\right),$$
$$\mathbb{V}(\hat{\tau}_f^{\delta}(\boldsymbol{Z})) = \mathbb{V}(\hat{\tau}_f(\boldsymbol{Z})) + \frac{n^2}{(n-1)^2}\delta p_f(1-p_f)\left(\frac{2\kappa_{11}}{n_1} + \frac{2\kappa_{00}}{n_0} + \delta\cdot\frac{n}{n_1 n_0}\right) + O\left(\frac{\delta}{n^2}\right),$$

where $\kappa_{st} = \mathbb{E}[Y_i(s) \mid f(\boldsymbol{X}_i) = t]$. *Minimizing these variances over* δ *results in the following optimal value of* δ_λ^* *and* δ_τ^* *for the PAV and PAPE, respectively*:

$$\delta_\lambda^* = -\left(\frac{n_0}{n}\kappa_{11} + \frac{n_1}{n}\kappa_{00}\right); \quad \delta_\tau^* = \delta_\lambda^* + O\left(\frac{1}{n}\right).$$

Proof is in Appendix B. The proposition implies that if we wish to minimize variance across a range of $f(\boldsymbol{X}_i)$, then when $n_1 = n_0 = n/2$, the optimal value of δ approximately balances the two potential outcomes around zero after a constant shift, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\{(Y_i(1) + \delta_\lambda^*) + (Y_i(0) + \delta_\lambda^*)\} \approx 0.$$

4 *Ex-ante* vs *ex-post* experimental evaluations

So far, we have considered an *ex-post* evaluation, in which we first conduct a completely randomized experiment and then evaluate ITRs using the data from the experiment. Alternatively, researchers may consider an *ex-ante* experimental evaluation, in which we randomly assign units to an ITR, i.e., the ITR itself is the "treatment" of this experiment. *Ex-ante* experimental designs are commonly used in practice [see 24,25, for example].

We apply the Neyman's repeated sampling framework to compare the statistical efficiency of *ex-ante* and *ex-post* experimental evaluations. We show below that, perhaps surprisingly, in some cases, *ex-post* evaluation is more efficient than *ex-ante* evaluation. Our result suggests that given a potential ethical concern of *ex-ante* experimental evaluation, researchers may prefer *ex-post* evaluation. Another reason to prefer *ex-post* evaluation is that this design allows one to evaluate any number of ITRs, while the *ex-ante* evaluation is tied to a particular ITR. In this section, our analysis focuses on the PAPE. Since the PAV does not compare between two different treatment regimes, it does not make sense to design a randomized trial around it.

4.1 Setup

For the *ex-ante* evaluation of the PAPE, we assume a simple random of n units from the same target population, $\mathcal{P}$. Consider a completely randomized experiment, in which a total of n_f units are randomly assigned to an ITR f, while the remaining units $n_r = n - n_f$ are assigned to the random treatment rule with the probability of treatment

assignment equal to n_{r1}/n_r. Let F_i be an indicator variable, which is equal to 1 if unit i is assigned to the ITR f and is equal to 0 otherwise. Under the random treatment rule, the number of units that are randomly assigned to the treatment condition is n_{r1}, while $n_{r0} = n_r - n_{r1}$ units are assigned to the control condition. As before, we use T_i to represent the treatment indicator. We formally state these assumptions.

Assumption 4. (Complete randomization in the *ex-ante* evaluation of PAPE) The probability of being assigned to the ITR rather than the random treatment rule is given by

$$\Pr(\boldsymbol{F} = \boldsymbol{f} \mid \{Y_i(1), Y_i(0), \boldsymbol{X}_i\}_{i=1}^n) = \frac{1}{\binom{n}{n_f}},$$

for each $\boldsymbol{f}$, where $\sum_{i=1}^n f_i = n_f$. Among those who are assigned to the random treatment rule, i.e., $F_i = 0$, the probability of treatment assignment is given by

$$\Pr(\boldsymbol{T} = \boldsymbol{t} \mid \{Y_i(1), Y_i(0), \boldsymbol{X}_i\}_{i=1}^n) = \frac{1}{\binom{n_r}{n_{r1}}},$$

for each $\boldsymbol{t}$ where $\sum_{i=1}^n (1 - F_i) t_i = n_{r1}$.

Using this experimental data, we wish to estimate the PAPE defined in equation (4). For simplicity, we have the number of treated units under the random treatment rule to equal that under the ITR condition, i.e., $\hat{p}_f = n_{r1}/n_r$, where $\hat{p}_f = \sum_{i=1}^n f(\boldsymbol{X}_i)/n$, so that the evaluation estimator would not need to be further adjusted. Fortunately, in practice, this can be easily accomplished so long as the covariates are available prior to the randomization of treatment assignment among the group assigned to the random treatment rule. Finally, the so-called Neyman allocation implies that if the variances of $Y_i(1)$ and $Y_i(0)$ differ significantly from one another, one can gain additional statistical efficiency by allocating more units to the treatment condition whose potential outcome has a greater variance. This optimal design, however, does require the availability of external data to estimate these variances. For the sake of simplicity, we do not consider such optimal designs here.

We consider the following estimator of the PAPE for the *ex-ante* experimental evaluation that accounts for a potential difference in the proportion of treated units between the ITR and the random treatment rule by appropriately weighting the latter:

$$\hat{\tau}_f^*(\boldsymbol{Z}_n) = \frac{n}{n-1}\left(\frac{1}{n_f}\sum_{i=1}^n F_i Y_i - \frac{\hat{p}_f}{n_{r1}}\sum_{i=1}^n (1 - F_i) T_i Y_i - \frac{1 - \hat{p}_f}{n_{r0}}\sum_{i=1}^n (1 - F_i)(1 - T_i) Y_i\right). \tag{8}$$

The *ex-ante* evaluation differs from the *ex-post* evaluation in two ways. First, the *ex-ante* estimator requires two separate random assignments (T_i and F_i), while the *ex-post* estimator only involves one. Intuitively, an additional layer of randomization increases variance. Second, the *ex-ante* evaluation requires a separate group that follows an ITR, whereas all individuals under the *ex-post* evaluation are simply randomly assigned either to the treatment or control group. As a result, under the *ex-post* evaluation, we use the samples identically, which could further reduce the variance. Together, we expect the *ex-ante* evaluation to be less efficient than the *ex-post* evaluation as the full sample is not utilized for every part of the estimation. In the following, we use Neyman's repeated sampling framework to confirm this intuition under a set of simplifying assumptions.

4.2 Comparison of the two experimental designs

Before comparing two modes of evaluation, we derive the bias and variance of the *ex-ante* evaluation estimators under the Neyman's repeated sampling framework. In the current case, the uncertainty comes from three types of randomness: (1) the random assignment to the individualized or random treatment rule, (2) the randomized treatment assignment under the random assignment rule, and (3) the simple random sampling of units from the target population. The next theorem shows that this estimator is unbiased and the variance is identifiable. Proof is given in Appendix A.

Theorem 3. (Unbiasedness and variance of the *ex-ante* PAPE estimator) *Under Assumptions* 1, 3, *and* 4, *the expectation and variance of the ex-ante PAPE estimator defined in equation* (8) *are given by*

$$\mathbb{E}(\hat{\tau}_f^*(\boldsymbol{Z}_n)) = \tau_f,$$
$$\mathbb{V}(\hat{\tau}_f^*(\boldsymbol{Z}_n)) = \frac{n^2}{(n-1)^2}\left[\mathbb{E}\left[\frac{S_f^2}{n_f} + \frac{\hat{p}_f^2 S_1^2}{n_{r1}} + \frac{(1-\hat{p}_f)^2 S_0^2}{n_{r0}}\right] + \frac{1}{n^2}\{\tau_f^2 - np_f(1-p_f)\tau^2 + 2(n-1)(2p_f-1)\tau_f\tau\}\right],$$

where $S_f^2 = \sum_{i=1}^n \{Y_i(f(\boldsymbol{X}_i)) - \overline{Y(f(\boldsymbol{X}))}\}^2/(n-1)$, *and* $S_t^2 = \sum_{i=1}^n (Y_i(t) - \overline{Y(t)})^2/(n-1)$ *with* $\overline{Y(f(\boldsymbol{X}))} = \sum_{i=1}^n Y_i(f(\boldsymbol{X}_i))/n$ *and* $\overline{Y(t)} = \sum_{i=1}^n Y_i(t)/n$ *for* $t = 0, 1$.

Given these results, we examine the relative statistical efficiency of the *ex-post* and *ex-ante* experimental evaluations. To facilitate the comparison, we assume $n_1 = n_0 = n_f = n_r = n/2$. In words, the *ex-post* evaluation sets the treatment assignment probability to 1/2, and the *ex-ante* evaluation also sets the probability of being assigned to the ITR to 1/2. In the same fashion, we also assume $n_{r1} = n_{r0} = n/4$, implying that the *ex-ante* evaluation sets the treatment assignment probability under the random treatment rule to 1/2 as well. Although our result below may not be applicable beyond this simplified setting, we believe that this equal allocation setting is a common choice in practice and therefore is worthy of investigation.

Under this simplified setting, the difference in the variance of the PAPE estimator between the *ex-ante* and *ex-post* evaluations is given by

$$\begin{aligned}
\mathbb{V}(\hat{\tau}_f^*(\boldsymbol{Z}_n)) - \mathbb{V}(\hat{\tau}_f(\boldsymbol{Z}_n)) &= \frac{2n}{(n-1)^2}\Big[\mathbb{E}\{p_f^2 S_1^2 + (1-p_f)^2 S_0^2\} + 2\mathrm{Cov}(f(\boldsymbol{X}_i)Y_i(1), (1-f(\boldsymbol{X}_i))Y_i(0)) \\
&\quad + 2p_f\mathrm{Cov}(f(\boldsymbol{X}_i)Y_i(1), Y_i(1)) + 2(1-p_f)\mathrm{Cov}((1-f(\boldsymbol{X}_i))Y_i(0), Y_i(0))\Big] \\
&= \frac{2n}{(n-1)^2}\Big[p_f^2\mathbb{V}(Y_i(1)) + (1-p_f)^2\mathbb{V}(Y_i(0)) - 2p_f(1-p_f)\mathbb{E}(Y_i(0)\mid f(\boldsymbol{X}_i)=0) \\
&\quad \times \mathbb{E}(Y_i(1)\mid f(\boldsymbol{X}_i)=1) \\
&\quad + 2p_f^2\{\mathbb{E}(Y_i^2(1)\mid f(\boldsymbol{X}_i)=1) - \mathbb{E}(Y_i(1))\mathbb{E}(Y_i(1)\mid f(\boldsymbol{X}_i)=1)\} \\
&\quad + 2(1-p_f)^2\{\mathbb{E}(Y_i^2(0)\mid f(\boldsymbol{X}_i)=0) - \mathbb{E}(Y_i(0))\mathbb{E}(Y_i(0)\mid f(\boldsymbol{X}_i)=0)\}\Big].
\end{aligned} \tag{9}$$

The details of the derivation are given in Appendix C. Suppose now that the ITR correctly assigns individuals on average, i.e., $\mathbb{E}(Y_i(t)\mid f(\boldsymbol{X}_i)=t) \geq \mathbb{E}(Y_i(t)\mid f(\boldsymbol{X}_i)=1-t)$ for $t = 0, 1$. Under this assumption, the last two terms in the square bracket are positive, i.e.,

$$\mathbb{E}(Y_i^2(t)\mid f(\boldsymbol{X}_i)=t) - \mathbb{E}(Y_i(t))\mathbb{E}(Y_i(t)\mid f(\boldsymbol{X}_i)=t) \geq \mathbb{V}(Y_i(t)\mid f(\boldsymbol{X}_i)=t),$$

for $t = 0, 1$. Hence, the only term that is possibly negative in equation (9) is the third term in the square bracket. For simplicity, further assume that we shift the outcomes to minimize variance of the *ex-post* estimator and achieved $\mathbb{E}(Y_i(1)+Y_i(0)\mid f(\boldsymbol{X}_i)=1) = \mathbb{E}(Y_i(1)+Y_i(0)\mid f(\boldsymbol{X}_i)=0) = 0$ (see equation (1)). This guarantees that the optimal choice of δ is zero, and hence, no adjustment in variance is necessary. Under this assumption, we can bound equation (9) from below as follows (see Appendix D for details):

$$\begin{aligned}
&\mathbb{V}(\hat{\tau}_f^*(\boldsymbol{Z}_n)) - \mathbb{V}(\hat{\tau}_f(\boldsymbol{Z}_n)) \\
&= \frac{2n}{(n-1)^2}\Big[p_f^2\mathbb{V}(Y_i(1)) + (1-p_f)^2\mathbb{V}(Y_i(0)) + 2p_f^2\mathbb{V}(Y_i(1)\mid f(\boldsymbol{X}_i)=1) + 2(1-p_f)^2\mathbb{V}(Y_i(0)\mid f(\boldsymbol{X}_i)=0) \\
&\quad + 2p_f(1-p_f)[(1-p_f)\{\mathbb{E}(Y_i(0)\mid f(\boldsymbol{X}_i)=0)\}^2 + p_f\{\mathbb{E}(Y_i(1)\mid f(\boldsymbol{X}_i)=1)\}^2]\Big] \geq 0.
\end{aligned}$$

The result implies that under a set of simplifying assumptions, the *ex-post* evaluation is more efficient than the *ex-ante* evaluation. We note, however, that this conclusion may not hold if the *ex-ante* and *ex-post* setups have sample allocation different from the setting considered here.

5 Incorporating the uncertainty of ML training

In the aforementioned sections, we assume that the ITR to be evaluated is given. For example, an ITR may be derived using an external dataset. But, in many cases, researchers may wish to use the same experimental dataset to both derive an ITR and evaluate it. One possibility is to randomly split a dataset into the training and evaluation datasets, and then use the former to learn an ITR and the latter for its evaluation. Unfortunately, this *sample splitting* approach does not use the data most efficiently.

An alternative and more efficient approach is *cross-fitting*. The idea is to randomly split the data into K folds of equal size and then use each fold as the evaluation data while using the remaining $K - 1$ folds as the training data to learn an ITR. By repeating this process across K folds and averaging the evaluation results, we are able to use the entire dataset for both training and evaluation.

While the dominant "double machine learning" (DML) approach uses the same cross-fitting procedure [26], we show here that Neyman's repeated sampling framework can also incorporate this cross-fitting approach. Unlike the DML, Neyman's framework enables us to derive the finite-sample properties of ITR evaluation solely based on the random splitting of the data as well as randomization of treatment assignment and random sampling of units.

5.1 Setup

Consider a generic ML algorithm, which we define as a deterministic function mapping the space of training data of finite size, denoted by $\mathcal{Z}$, to the space of all possible scoring rules $\mathcal{S}$,

$$F : \mathcal{Z} \to \mathcal{S}. \tag{10}$$

Typically, the scoring rule of interest is the estimated CATE such that the largest value indicates the highest treatment prioritization. Alternatively, the scoring rule may be based on the estimated baseline risk, i.e., $\mathbb{E}(Y_i(0)|\boldsymbol{X}_i = \boldsymbol{x})$. We do not, however, assume that the ML algorithm used to generate the scoring rule accurately estimates either the CATE or baseline risk. Indeed, we essentially impose no assumption on how the scoring rule is created. Once the scoring rule is estimated by an ML algorithm, the ITR is given by

$$\hat{f}_{\boldsymbol{Z}_n}(\boldsymbol{x}) \coloneqq \mathbf{1}\{F(\boldsymbol{Z}_n)(\boldsymbol{x}) > 0\}, \tag{11}$$

where the notation makes it explicit that the ITR depends on the specific training data $\boldsymbol{Z}_n \in \mathcal{Z}$ of sample size n.

Next, consider the following standard cross-fitting procedure. First, we randomly split the experimental data of size n into K subsamples of equal size $m = n/K$, where, for notational simplicity, we assume n is a multiple of K. Then, for each $k = 1, 2, \ldots, K$, we use the kth subsample as an evaluation dataset $\boldsymbol{Z}_m^{(k)} = \{\boldsymbol{X}_i^{(k)}, T_i^{(k)}, Y_i^{(k)}\}_{i=1}^{m}$, while the remaining $(K - 1)$ subsamples are used as the training dataset $\boldsymbol{Z}_{n-m}^{(-k)} = \{\boldsymbol{X}_i^{(-k)}, T_i^{(-k)}, Y_i^{(-k)}\}_{i=1}^{n-m}$. Without loss of generality, we assume that the number of treated (control) units is identical across K folds and denote it using m_1 ($m_0 = m - m_1$).

Then, for each fold k, we estimate an ITR by applying the ML algorithm F to the training data $\boldsymbol{Z}_{n-m}^{(-k)}$, which we denote by $\hat{f}^{(-k)} = \hat{f}_{\boldsymbol{Z}_{n-m}^{(-k)}}$. We then evaluate the performance of the ML algorithm F by computing an evaluation metric of interest based on the test data $\boldsymbol{Z}_m^{(k)}$. Repeating this process K times for each k and averaging the results gives a cross-fitting estimator of the evaluation metric. Here, we focus on the cross-fitting PAV estimator:

$$\hat{\lambda}_K^F(\boldsymbol{Z}_n) = \frac{1}{K}\sum_{k=1}^{K} \hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}_m^{(k)}), \tag{12}$$

where $\hat{\lambda}_f(\cdot)$ is defined in equation (6). We now discuss the estimand, for which this cross-fitting estimator is unbiased.

5.2 Evaluation metrics under cross-fitting

To extend Neyman's repeated sampling framework to cross-fitting with $K \geq 2$ folds, we begin by noting that the ITR in this setting varies as a function of training data. Thus, we consider the performance measure that averages over the random sampling of training data as well as the randomization of treatment assignment and random sampling of units. In other words, we evaluate the average performance of ITR that is generated by the application of ML algorithm F across K different (but overlapping) training datasets. This contrasts with the performance evaluation metric of a fixed ITR discussed in earlier sections.

For the PAV under cross-fitting, we consider an average ITR over across training data of size $n - m$:

$$\bar{f}^F_{n-m}(\boldsymbol{X}_i) = \mathbb{E}_{\boldsymbol{Z}_{n-m}}\{\hat{f}_{\boldsymbol{Z}_{n-m}}(\boldsymbol{X}_i) \mid \boldsymbol{X}_i\} = \mathbb{P}_{\boldsymbol{Z}_{n-m}}\{\hat{f}_{\boldsymbol{Z}_{n-m}}(\boldsymbol{X}_i) = 1 \mid \boldsymbol{X}_i\},$$

which represents the proportion of times the estimated ITR would assign the treatment to a unit with a specific value of covariates. The notation makes explicit the dependence on the size of training data $n - m$ as well as the ML algorithm F.

Under Neyman's repeated sampling framework, one can view each estimated ITR as another random sampling from a population of ITRs based on the ML algorithm F with training dataset of size $n - m$. Thus, the PAV under cross-fitting can be defined as

$$\lambda^F_{n-m} \coloneqq \mathbb{E}\{\bar{f}^F_{n-m}(\boldsymbol{X}_i)Y_i(1) + (1 - \bar{f}^F_{n-m}(\boldsymbol{X}_i))Y_i(0)\}.$$

For the PAPE, we consider the cross-fitting version of the proportion treated by ITR p_f as follows:

$$p^F_{n-m} \coloneqq \mathbb{P}_{\boldsymbol{Z}_{n-m}}\{\hat{f}_{\boldsymbol{Z}_{n-m}}(\boldsymbol{X}_i) = 1\}.$$

Then, the PAPE under cross-fitting can be defined as

$$\tau^F_{n-m} \coloneqq \mathbb{E}\{\bar{f}^F_{n-m}(\boldsymbol{X}_i)Y_i(1) + (1 - \bar{f}^F_{n-m}(\boldsymbol{X}_i))Y_i(0) - p^F_{n-m}Y_i(1) - (1 - p^F_{n-m})Y_i(0)\}. \tag{13}$$

As shown earlier, the PAPE is equal to the covariance between the average proportion treated and the individual treatment effect:

$$\tau^F_{n-m} = \mathrm{Cov}(\bar{f}^F_{n-m}(\boldsymbol{X}_i), Y_i(1) - Y_i(0)).$$

5.3 Finite sample properties

We now apply Neyman's repeated sampling framework to the cross-fitting PAV estimator given in equation (12). It is easy to show that $\hat{\lambda}^F_K$ is an unbiased estimator of λ^F_{n-m}. To derive the variance, we first note that the evaluation metric is correlated across K folds because cross-fitting utilizes each subsample for both training and testing:

$$\mathbb{V}(\hat{\lambda}^F_K(\boldsymbol{Z}_n)) = \frac{\mathbb{V}(\hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}^{(k)}_m))}{K} + \frac{K-1}{K}\mathrm{Cov}(\hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}^{(k)}_m), \hat{\lambda}_{\hat{f}^{(-\ell)}}(\boldsymbol{Z}^{(\ell)}_m)),$$

where $k \neq \ell$. We then use a useful lemma about cross-fitting due to [27] and rewrite the covariance term as follows:

$$\mathrm{Cov}(\hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}^{(k)}_m), \hat{\lambda}_{\hat{f}^{(-\ell)}}(\boldsymbol{Z}^{(\ell)}_m)) = \mathbb{V}(\hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}^{(k)}_m)) - \mathbb{E}(S^2_F),$$

where S^2_F is the sample variance of $\hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}^{(k)}_m)$ across K folds. Putting them together, we have,

$$\mathbb{V}(\hat{\lambda}^F_K(\boldsymbol{Z}_n)) = \mathbb{V}(\hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}^{(k)}_m)) - \frac{K-1}{K}\mathbb{E}(S^2_F). \tag{14}$$

We can further analyze the first term of equation (14) by following the analytical strategy used in Theorem 1. The only difference is that the estimated ITR is correlated across observations due to training process:

$$\mathbb{V}(\hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}^{(k)}_m)) = \frac{\mathbb{E}(S^2_{\hat{f}1})}{m_1} + \frac{\mathbb{E}(S^2_{\hat{f}0})}{m_0} + \mathrm{Cov}(Y_{\hat{f}i}(1) - Y_{\hat{f}i}(0), Y_{\hat{f}j}(1) - Y_{\hat{f}j}(0)), \tag{15}$$

where $i \neq j$, $Y_{\hat{f}i}(t) = \mathbf{1}\{\hat{f}^{(-k)}(\boldsymbol{X}_i) = t\}Y_i(t)$, and $S^2_{\hat{f}t}$ is the sample variance of $Y_{\hat{f}i}(t)$. Further simplifying the covariance term yields the following theorem whose proof is given in Appendix A.5.1. of our previously published work [14].

Theorem 4. (Unbiasedness and exact variance of the cross-fitting PAV estimator [14]) *Under Assumptions* 1–3 *the expectation and variance of the cross-fitting PAV estimator defined in equation* (12) *are given by,*

$$\mathbb{E}(\hat{\lambda}^F_K(\boldsymbol{Z}_n)) = \lambda^F_{n-m},$$
$$\mathbb{V}(\hat{\lambda}^F_K(\boldsymbol{Z}_n)) = \frac{\mathbb{E}(S^2_{\hat{f}1})}{m_1} + \frac{\mathbb{E}(S^2_{\hat{f}0})}{m_0} + \mathbb{E}\{\mathrm{Cov}(\hat{f}^{(-k)}(\boldsymbol{X}_i), \hat{f}^{(-k)}(\boldsymbol{X}_j) \mid \boldsymbol{X}_i, \boldsymbol{X}_j)\tau_i\tau_j\} - \frac{K-1}{K}\mathbb{E}(S^2_F),$$

for $i \neq j$, *where* $\tau_i = Y_i(1) - Y_i(0)$, $S^2_{\hat{f}t} = \sum_{i=1}^m (Y_{\hat{f}i}(t) - \overline{Y_{\hat{f}}(t)})^2/(m-1)$, $S^2_F = \sum_{k=1}^K (\hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}^{(k)}_m)) - \overline{\hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}^{(k)}_m)}^2/(K-1)$ *with* $Y_{\hat{f}i}(t) = \mathbf{1}\{\hat{f}^{(-k)}(\boldsymbol{X}_i) = 1\}Y_i(t)$, $\overline{Y_{\hat{f}}(t)} = \sum_{i=1}^m Y_{\hat{f}i}(t)/m$, *and* $\overline{\hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}^{(k)}_m)} = \sum_{k=1}^K \hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}^{(k)}_m)/K$, *for* $t = \{0, 1\}$.

In particular, we note that when compared with the fixed ITR setting, there are two additional terms. One of these terms is proportional to $\mathrm{Cov}(\hat{f}^{(-k)}(\boldsymbol{X}_i), \hat{f}^{(-k)}(\boldsymbol{X}_j) \mid \boldsymbol{X}_i, \boldsymbol{X}_j)$, which represents the covariance between the evaluation samples due to the training process, and the product of individual treatment effects $\tau_i\tau_j$. This term is often positive because if the ITR is estimated well, it is more likely to make the same treatment assignment to units when their individual treatment effects are similar. In numerical experiments, we typically find that this term is usually relatively small. The other term $-\frac{K-1}{K}\mathbb{E}(S^2_F)$ is always negative and quantifies the efficiency gain resulting from utilizing the cross-validation procedure. Furthermore, from Lemma 1 in [27], we can show that

$$\mathbb{V}(\hat{\lambda}_{\hat{f}^{(-k)}}(\boldsymbol{Z}^{(k)}_m)) = \frac{\mathbb{E}(S^2_{\hat{f}1})}{m_1} + \frac{\mathbb{E}(S^2_{\hat{f}0})}{m_0} + \mathbb{E}\{\mathrm{Cov}(\hat{f}^{(-k)}(\boldsymbol{X}_i), \hat{f}^{(-k)}(\boldsymbol{X}_j) \mid \boldsymbol{X}_i, \boldsymbol{X}_j)\tau_i\tau_j\} \geq \mathbb{E}(S^2_F).$$

Therefore, maximally, the efficiency gain resulting from the cross-fitting procedure reduces the variance to 1/K $\mathbb{E}(S^2_F)$ when the estimated PAV from each of K folds is completely independent.

6 A numerical study

In this section, we empirically validate our theoretical results through a numerical study. In particular, we focus on demonstrating the results related to the lack of invariance (Proposition 1) and the efficiency comparison between the *ex-ante* and *ex-post* estimators (Theorem 3). Strong finite-sample performance of the proposed estimators have been extensively demonstrated in our previously published study [14].

In all our simulations, we use the 28th data-generating process (DGP) from the 2016 Atlantic Causal Inference Conference (ACIC) Competition, of which the details are given in [28]. For the population distribution of pre-treatment covariates, we use the empirical distribution of covariates from this sample of $n = 4{,}802$ observations with 58 covariates $\boldsymbol{X}$, including 3 categorical, 5 binary, 27 count data, and 13 continuous variables, i.e., we obtain each simulation sample via bootstrap. We further assume that the treatment assignment is completely randomized, and the treatment and control groups are of equal size, i.e., $n_1 = n_0 = n/2$. Finally, the formula for the outcome model is reproduced in Appendix E.

First, we investigate the effect of shifting potential outcomes by a constant on the variance of estimators. Figure 2(a) plots the empirical standard deviation of the PAV estimator (the vertical axis) with $n = 100$ samples from the DGP as a function of constant shift in potential outcomes (the horizontal axis). Here, we centered the potential outcomes shift so that the optimal value of δ given in Proposition 1 is zero, i.e.,

$$\frac{n_0}{n}\kappa_{11} + \frac{n_1}{n}\kappa_{00} = 0.$$

As predicted by our theoretical analysis, we find that balancing the potential outcomes leads to a lower standard error in the estimator due to the unbalanced nature of the relevant potential outcomes $\mathbf{1}\{f(\boldsymbol{X}_i) = t\}Y_i(t)$.

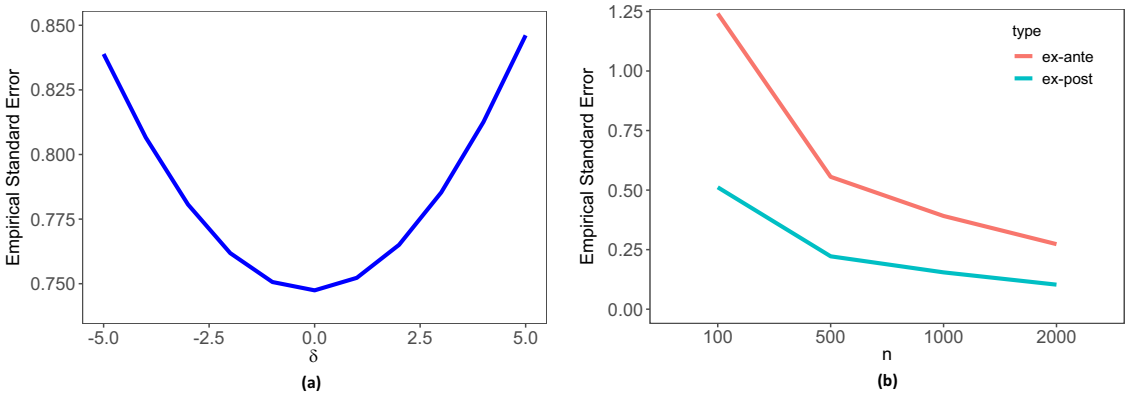

Figure 2: Numerical experiments: (a) Empirical standard error of PAV estimator as a function of constant shift in potential outcomes. $\delta = 0$ minimizes the standard error of PAV and (b) comparison of empirical standard error of the *ex-ante* and *ex-post* PAPE estimators (y-axis) for various sample sizes (x-axis).

Second, we compare the statistical efficiency of the *ex-ante* and *ex-post* PAPE estimators under the assumption $n_f = n_r = n/2$ and $n_{r1} = n_{r0} = n/4$. Consistent with our theoretical results, Figure 2(b) shows that the standard error of the *ex-ante* estimator is consistently greater than that of the *ex-post* estimator. For example, when the sample size is 500, the former is over twice the latter.

7 Conclusion

In this article, we provided a short overview of how Neyman's repeated sampling framework can be used to experimentally evaluate the performance of arbitrary ITRs. We consider the two settings, one in which an ITR is given and the other in which an ITR is estimated from the same data. We also demonstrated the new challenges that result from the application of Neyman's framework, including the lack of invariance of evaluation estimators and the need to incorporate the uncertainty due to training of ML algorithms. We further demonstrated how Neyman's repeated-sampling framework can highlight the difference between the *ex-ante* evaluation and *ex-post* evaluation of ITRs by showing that the *ex-post* evaluation is statistically more efficient. Our ongoing work also applies this framework to the estimation of heterogeneous treatment effects discovered by ML algorithms [29]. Altogether, we have shown that a century after his original proposal, Neyman's analytical framework remains relevant and is widely applicable to the evaluation of today's causal ML methods.

Acknowledgement: The authors would like to thank Peng Ding and the two anonymous reviewers for their helpful and invaluable feedback during the review process.

Funding information: None declared.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: Prof. Kosuke Imai is a member of the Editorial Advisory Board in the Journal of Causal Inference but was not involved in the review process of this article.

Data availability statement: The numerical experiments included in the current study can be reproduced with the R scripts available at https://github.com/MichaelLLi/NeymanMLCode.

References

[1] Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Ann Agricultural Sci. 1923:1–51.
[2] Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. Ann Appl Stat. 2013;7:443–70.
[3] Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proc Nat Acad Sci. 2016;113(27):7353–60.
[4] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Amer Stat Assoc. 2018;113(523):1228–42.
[5] Hahn PR, Murray JS, Carvalho CM. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. Bayesian Anal. 2020;15(3):965–1056.
[6] Dudík M, Langford J, Li L. Doubly robust policy evaluation and learning. in Proceedings of the 28th International Conference on International Conference on Machine Learning. ICML'11, USA: Omnipress; 2011. p. 1097–104.
[7] Zhang B, Tsiatis AA, Davidian M, Laber E. Estimating optimal treatment regimes from a classification perspective. Stat. 2012;1(1):103–14.
[8] Chakraborty B, Laber E, Zhao Y-Q. Inference about the expected performance of a data-driven dynamic treatment regime. Clin Trials. 2014;11(4):408–17.
[9] Jiang N, Li L. Doubly robust off-policy value evaluation for reinforcement learning. in: Proceedings of The 33rd International Conference on Machine Learning. Balcan MF, Weinberger KQ, (Eds.), vol. 48 of Proceedings of Research. New York, New York, USA: PMLR; 20–22 Jun 2016. p. 652–61.
[10] Kallus N. Balanced policy evaluation and learning. in: Advances in Neural Information Processing Systems 31. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R. (Eds.) Curran Associates, Inc.; 2018. p. 8895–906.
[11] Qi Z, Liu D, Fu H, Liu Y. Multi-armed angle-based direct learning for estimating optimal individualized treatment rules with various outcomes. J Amer Stat Assoc. 2020;115(530):678–91.
[12] Mo W, Liu Y. Efficient learning of optimal individualized treatment rules for heteroscedastic or misspecified treatment-free effect models. J R Stat Soc Ser B Stat Methodol. 2022;84(2):440–72.
[13] Ben-Michael E, Greiner J, Imai K, Jiang Z. Safe policy learning through extrapolation: Application to pre-trial risk assessment. Technical Report. 2021. arXiv:2109.11679.
[14] Imai K, Liii ML. Experimental evaluation of individualized treatment rules. J Amer Stat Assoc. 2023;118(541):242–56.
[15] Rubin DB. Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9 by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T.P. Speed". Stat Sci. 1990;5:472–80.
[16] Ding P, Li X, Miratrix LW. Bridging finite and super population causal inference. J Causal Inference. 2017;5(2):20160027.
[17] Qian M, Murphy SA. Performance gurantees for individualized treatment rules. Ann Stat. 2011;39(2):1180–210.
[18] Luedtke AR, van der Laan MJ. Optimal individualized treatments in resource-limited settings. Int J Biostat. 2016;12(1):283–303.
[19] Luedtke AR, van der Laan MJ. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. Ann Statist. 2016;44(2):713–42.
[20] Zhou X, Mayer-Hamblett N, Khan U, Kosorok MR. Residual weighted learning for estimating individualized treatment rules. J Amer Stat Assoc. 2017;112(517):169–87.
[21] Kitagawa T, Tetenov A. Who should be treated?: Empirical welfare maximization methods for treatment choice. Econometrica 2018;86:591–616.
[22] Radcliffe NJ. Using control groups to target on predicted lift: Building and assessing uplift models. Direct Market Analytic J. 2007;1(3):14–21.
[23] Yadlowsky S, Fleming S, Shah N, Brunskill E, Wager S. Evaluating treatment prioritization rules via rank-weighted average treatment effects. 2021. arXiv: http://arXiv.org/abs/arXiv:211107966.
[24] Kumar A, Aikens RC, Hom J, Shieh L, Chiang J, Morales D, et al. Orderrex clinical user testing: a randomized trial of recommender system decision support on simulated cases. J Amer Med Inform Assoc. 2020;27(12):1850–9.
[25] Forman EM, Goldstein SP, Crochiere RJ, Butryn ML, Juarascio AS, Zhang F, et al. Randomized controlled trial of ontrack, a just-in-time adaptive intervention designed to enhance weight loss. Translat Behav Med. 2019;9(6):989–1001.
[26] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. Oxford, UK: Oxford University Press; 2018.
[27] Nadeau C, Bengio Y. Inference for the generalization error. Machine Learning. 2003;52(3):239–81.
[28] Dorie V, Hill J, Shalit U, Scott M, Cervone D, Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. Stat Sci. Vol. 34. February 2019; p. 43–68.
[29] Imai K, Li ML. Statistical inference for heterogeneous treatment effects discovered by generic machine learning in randomized experiments. Journal of Business & Economic Statistics. Forthcoming.
[30] Neyman J. On the application of probability theory to agricultural experiments: Essay on principles, section 9 (translated in 1990). Stat Sci. 1923;5:465–80.

Appendix

A Proof of Theorem 3

We first consider the following intermediate estimator:

$$\tilde{\tau}_f^*(\boldsymbol{Z}_n) = \frac{1}{n_f}\sum_{i=1}^{n} Y_i(f(\boldsymbol{X}_i))F_i - \frac{1}{n_r}\sum_{i=1}^{n} Y_i(T_i)(1-F_i). \tag{A1}$$

This estimator differs from the *ex-ante* estimator of the PAPE $\hat{\tau}_f^*$, by a small factor, i.e., $\tilde{\tau}_f^* = (n-1)/n\hat{\tau}_f$ under the condition that $\hat{p}_f = n_{r1}/r_1$. The following lemma derives the expectation and variance of this estimator. Using this lemma, the results of Theorem 3 can be obtained immediately.

Lemma 1. (Expectation and variance of the intermediate estimator) *Under Assumptions* 1, 3, *and* 4, *the expectation and variance of the estimator given in equation* (A1) *for estimating the PAPE defined in equation* (4) *are given by*:

$$\mathbb{E}(\tilde{\tau}_f^*(\boldsymbol{Z}_n)) = \frac{n-1}{n}\tau_f,$$
$$\mathbb{V}(\tilde{\tau}_f^*(\boldsymbol{Z}_n)) = \frac{\mathbb{E}(S_f^2)}{n_f} + \mathbb{E}\left[\frac{\hat{p}_f^2 S_1^2}{n_{r1}} + \frac{(1-\hat{p}_f)^2 S_0^2}{n_{r0}}\right] + \frac{1}{n^2}\{\tau_f^2 - np_f(1-p_f)\tau^2 + 2(n-1)(2p_f-1)\tau_f\tau\}.$$

Proof. We first derive the bias expression. First, we take the expectation with respect to T_i,

$$\begin{aligned}
&\mathbb{E}[\tilde{\tau}_f^*(\boldsymbol{Z}_n) \mid \{\boldsymbol{X}_i, Y_i(1), Y_i(0), F_i\}_{i=1}^n] \\
&= \mathbb{E}\left[\frac{1}{n_f}\sum_{i=1}^{n} Y_i(f(\boldsymbol{X}_i))F_i - \frac{1}{n_r}\sum_{i=1}^{n}\{Y_i(1)T_i + Y_i(0)(1-T_i)\}(1-F_i) \mid \{\boldsymbol{X}_i, Y_i(1), Y_i(0), F_i\}_{i=1}^n\right] \\
&= \frac{1}{n_f}\sum_{i=1}^{n} Y_i(f(\boldsymbol{X}_i))F_i - \frac{1}{n_r}\sum_{i=1}^{n}\left\{Y_i(1)\frac{\sum_{i=1}^n f(\boldsymbol{X}_i)}{n} + Y_i(0)\left(1 - \frac{\sum_{i=1}^n f(\boldsymbol{X}_i)}{n}\right)\right\}(1-F_i).
\end{aligned}$$

Next, we take the expectation with respect to F_i:

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n_f}\sum_{i=1}^{n} Y_i(f(\boldsymbol{X}_i))F_i - \frac{1}{n_r}\sum_{i=1}^{n}\left\{Y_i(1)\frac{\sum_{i=1}^n f(\boldsymbol{X}_i)}{n} + Y_i(0)\left(1 - \frac{\sum_{i=1}^n f(\boldsymbol{X}_i)}{n}\right)\right\}(1-F_i) \mid \{\boldsymbol{X}_i, Y_i(1), Y_i(0)\}_{i=1}^n\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i(f(\boldsymbol{X}_i)) - \frac{1}{n_r n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[Y_i(1)f(\boldsymbol{X}_j)(1-F_i) + Y_i(0)(1-f(\boldsymbol{X}_j))(1-F_i) \mid \{\boldsymbol{X}_i, Y_i(1), Y_i(0)\}_{i=1}^n] \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i(f(\boldsymbol{X}_i)) - \frac{1}{n_r n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{n_r n}{n^2}Y_i(1)f(\boldsymbol{X}_j) + Y_i(0)\frac{n_r n}{n^2}(1-f(\boldsymbol{X}_j))\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i(f(\boldsymbol{X}_i)) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(Y_i(1)f(\boldsymbol{X}_j) + Y_i(0)(1-f(\boldsymbol{X}_j))).
\end{aligned}$$

Finally, we take the expectation over the sampling of $\{\boldsymbol{X}_i, Y_i(1), Y_i(0)\}$:

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i(f(\boldsymbol{X}_i)) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{Y_i(1)f(\boldsymbol{X}_j) + Y_i(0)(1-f(\boldsymbol{X}_j))\}\right] \\
&= \mathbb{E}\{Y_i(f(\boldsymbol{X}_i))\} - p_f\mathbb{E}(Y_i(1)) - (1-p_f)\mathbb{E}(Y_i(0)) - \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\{\mathrm{Cov}(Y_i(1), f(\boldsymbol{X}_i)) + \mathrm{Cov}(Y_i(0), 1-f(\boldsymbol{X}_i))\} \\
&= \tau_f - \frac{1}{n}\mathrm{Cov}(Y_i(1) - Y_i(0), f(\boldsymbol{X}_i)) = \frac{n-1}{n}\tau_f.
\end{aligned}$$

For the variance expression, we proceed as follows:

$$\begin{aligned}\mathbb{V}(\tilde{\tau}_f^*(\boldsymbol{Z}_n)) &= \mathbb{V}\{\mathbb{E}(\tau_f^*(\boldsymbol{Z}_n) \mid \{\boldsymbol{X}_i, Y_i(1), Y_i(0), F_i\}_{i=1}^n)\} + \mathbb{E}\{\mathbb{V}(\tau_f^*(\boldsymbol{Z}_n) \mid \boldsymbol{X}_i, Y_i(1), Y_i(0), F_i)\} \\ &= \mathbb{V}\left[\frac{1}{n_f}\sum_{i=1}^n Y_i(f(\boldsymbol{X}_i))F_i - \frac{1}{n_r}\sum_{i=1}^n \{Y_i(1)\hat{p}_f + Y_i(0)(1-\hat{p}_f)\}(1-F_i)\right] \\ &\quad + \mathbb{E}\left\{\frac{1}{n_r^2}\mathbb{V}\left[\sum_{i=1}^n \{Y_iT_i + Y_i(1-T_i)\}(1-F_i) \mid \{\boldsymbol{X}_i, Y_i(1), Y_i(0), F_i\}_{i=1}^n\right]\right\}.\end{aligned}$$

For the first term, we further use the law of total variance by conditioning on the sample, and center F_i via the transformation $D_i = F_i - n_f/n$. For the second term, we use the results of [30], with the following notation:

$$S_t^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i(t) - \overline{Y(t)})^2, \quad S_{01} = \frac{1}{n-1}\sum_{i=1}^n (Y_i(0) - \overline{Y(0)})(Y_i(1) - \overline{Y(1)}),$$

for $t = 0, 1$. Then, the variance becomes

$$\begin{aligned}\mathbb{V}(\tilde{\tau}_f^*(\boldsymbol{Z}_n)) &= \mathbb{E}\left[\mathbb{V}\left\{\frac{1}{n}\sum_{i=1}^n D_i\left(\frac{n}{n_f}Y(f(\boldsymbol{X}_i)) + \frac{n}{n_r}\widehat{Y}_i\right) \mid \{\boldsymbol{X}_i, Y_i(1), Y_i(0)\}_{i=1}^n\right\}\right] \\ &\quad + \mathbb{V}\left\{\frac{1}{n}\sum_{i=1}^n (Y(f(\boldsymbol{X}_i)) - \widehat{Y}_i)\right\} + \mathbb{E}\left[\frac{1}{n_r}\left\{\frac{\hat{p}_f^2 n_{r0}S_1^2}{n_{r1}} + \frac{(1-\hat{p}_f)^2 n_{r1}S_0^2}{n_{r0}} - 2\hat{p}_f(1-\hat{p}_f)S_{01}\right\}\right],\end{aligned}$$

where $\widehat{Y}_i = \hat{p}_f Y_i(1) + (1-\hat{p}_f)Y_i(0)$. Then, we have

$$\begin{aligned}\mathbb{V}(\tilde{\tau}_f^*(\boldsymbol{Z}_n)) &= \frac{\mathbb{E}(S_f^2)}{n_f} + \frac{\mathbb{E}(S_m^2)}{n_r} + \mathbb{E}\left[\frac{1}{n_r}\left\{\frac{\hat{p}_f^2 n_{r0}S_1^2}{n_{r1}} + \frac{(1-\hat{p}_f)^2 n_{r1}S_0^2}{n_{r0}} - 2\hat{p}_f(1-\hat{p}_f)S_{01}\right\}\right] \\ &\quad + \mathrm{Cov}((f(\boldsymbol{X}_i) - \hat{p}_f)Y_i(1) - (f(\boldsymbol{X}_i) - \hat{p}_f)Y_i(0), (f(\boldsymbol{X}_i) - \hat{p}_f)Y_i(1) - (f(\boldsymbol{X}_i) - \hat{p}_f)Y_i(0)) \\ &= \frac{\mathbb{E}(S_f^2)}{n_f} + \frac{\mathbb{E}(S_m^2)}{n_r} + \mathbb{E}\left[\frac{1}{n_r}\left\{\frac{\hat{p}_f^2 n_{r0}S_1^2}{n_{r1}} + \frac{(1-\hat{p}_f)^2 n_{r1}S_0^2}{n_{r0}} - 2\hat{p}_f(1-\hat{p}_f)S_{01}\right\}\right] \\ &\quad + \frac{1}{n^2}\{\tau_f^2 - np_f(1-p_f)\tau^2 + 2(n-1)(2p_f-1)\tau_f\tau\} \\ &= \frac{\mathbb{E}(S_f^2)}{n_f} + \mathbb{E}\left\{\frac{\hat{p}_f^2 S_1^2}{n_{r1}} + \frac{(1-\hat{p}_f)^2 S_0^2}{n_{r0}}\right\} + \frac{1}{n^2}\{\tau_f^2 - np_f(1-p_f)\tau^2 + 2(n-1)(2p_f-1)\tau_f\tau\},\end{aligned}$$

where $S_m^2 = \frac{1}{n-1}\sum_{i=1}^n (\widehat{Y}_i - \overline{\widehat{Y}})^2$ and the last equality follows from $\mathbb{E}(S_m^2) = \mathbb{E}\{\hat{p}_f^2 S_1^2 + (1-\hat{p}_f^2)S_0^2 + 2\hat{p}_f(1-\hat{p}_f)S_{01}\}$. □

B Proof of Proposition 1

By definition, we have

$$
\begin{aligned}
&\mathbb{V}(\hat{\lambda}_f^{\delta}(\boldsymbol{Z})) - \mathbb{V}(\hat{\lambda}_f(\boldsymbol{Z})) \\
&= \mathbb{V}\left(\hat{\lambda}_f(\boldsymbol{Z}) + \frac{\delta}{n_1}\sum_{i=1}^{n} T_i f(\boldsymbol{X}_i) + \frac{\delta}{n_0}\sum_{i=1}^{n}(1-T_i)(1-f(\boldsymbol{X}_i))\right) - \mathbb{V}(\hat{\lambda}_f(\boldsymbol{Z})) \\
&= 2\delta\cdot\mathrm{Cov}\left(\hat{\lambda}_f(\boldsymbol{Z}), \frac{1}{n_1}\sum_{i=1}^{n} T_i f(\boldsymbol{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)(1-f(\boldsymbol{X}_i))\right) \\
&\quad + \delta^2\cdot\mathbb{V}\left(\frac{1}{n_1}\sum_{i=1}^{n} T_i f(\boldsymbol{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)(1-f(\boldsymbol{X}_i))\right) \\
&= 2\delta\left\{\frac{\mathrm{Cov}(f(\boldsymbol{X}_i)Y_i(1), f(\boldsymbol{X}_i))}{n_1} + \frac{\mathrm{Cov}((1-f(\boldsymbol{X}_i))Y_i(1), 1-f(\boldsymbol{X}_i))}{n_0}\right\} + \delta^2\left\{\frac{p_f(1-p_f)}{n_1} + \frac{p_f(1-p_f)}{n_0}\right\} \\
&= \delta p_f(1-p_f)\left(\frac{2}{n_1}\mathbb{E}[Y_i(1)\,|\,f(\boldsymbol{X}_i)=1] + \frac{2}{n_0}\mathbb{E}[Y_i(0)\,|\,f(\boldsymbol{X}_i)=0] + \delta\cdot\frac{n}{n_1 n_0}\right).
\end{aligned}
$$

Define $b = \frac{n}{n-1} = 1 + \frac{1}{n-1}$. Then, we have

$$
\begin{aligned}
&\mathbb{V}(\hat{\tau}_f^{\delta}(\boldsymbol{Z})) - \mathbb{V}(\hat{\tau}_f(\boldsymbol{Z})) \\
&= \mathbb{V}\left(\hat{\tau}_f(\boldsymbol{Z}) + \frac{b\delta}{n_1}\sum_{i=1}^{n} T_i f(\boldsymbol{X}_i) + \frac{b\delta}{n_0}\sum_{i=1}^{n}(1-T_i)(1-f(\boldsymbol{X}_i))\right) - \mathbb{V}(\hat{\tau}_f(\boldsymbol{Z})) \\
&= 2b\delta\cdot\mathrm{Cov}\left(\hat{\tau}_f(\boldsymbol{Z}), \frac{1}{n_1}\sum_{i=1}^{n} T_i f(\boldsymbol{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)(1-f(\boldsymbol{X}_i))\right) \\
&\quad + b^2\delta^2\cdot\mathbb{V}\left(\frac{1}{n_1}\sum_{i=1}^{n} T_i f(\boldsymbol{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)(1-f(\boldsymbol{X}_i))\right) \\
&= 2b^2\delta\cdot\mathrm{Cov}\left(\hat{\lambda}_f(\boldsymbol{Z}), \frac{1}{n_1}\sum_{i=1}^{n} T_i f(\boldsymbol{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)(1-f(\boldsymbol{X}_i))\right) \\
&\quad - 2b^2\delta\cdot\mathrm{Cov}\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), \frac{1}{n_1}\sum_{i=1}^{n} T_i f(\boldsymbol{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)(1-f(\boldsymbol{X}_i))\right) \\
&\quad + b^2\delta^2\cdot\mathbb{V}\left(\frac{1}{n_1}\sum_{i=1}^{n} T_i f(\boldsymbol{X}_i) + \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)(1-f(\boldsymbol{X}_i))\right) \\
&= b^2\delta p_f(1-p_f)\left(\frac{2}{n_1}\mathbb{E}[Y_i(1)\,|\,f(\boldsymbol{X}_i)=1] + \frac{2}{n_0}\mathbb{E}[Y_i(0)\,|\,f(\boldsymbol{X}_i)=0] + \delta\cdot\frac{n}{n_1 n_0}\right) \\
&\quad - 2b^2\delta\left\{\frac{n\mathbb{E}[(\hat{p}_f Y_i(1) - \hat{p}_f\overline{Y_i(1)})(f(\boldsymbol{X}_i)-\hat{p}_f)]}{n_1(n-1)} + \frac{n\mathbb{E}[\{(1-\hat{p}_f)Y_i(0) - (1-\hat{p}_f)\overline{Y_i(0)}\}\{1-f(\boldsymbol{X}_i)-(1-\hat{p}_f)\}]}{n_0(n-1)}\right) \\
&= b^2\delta p_f(1-p_f)\left(\frac{2}{n_1}\mathbb{E}[Y_i(1)\,|\,f(\boldsymbol{X}_i)=1] + \frac{2}{n_0}\mathbb{E}[Y_i(0)\,|\,f(\boldsymbol{X}_i)=0] + \delta\cdot\frac{n}{n_1 n_0}\right) \\
&\quad - 2b^2\delta\left\{\frac{p_f + p_f^2(n-2)}{n_1 n^2}(\kappa_{11} - \mathbb{E}[Y_i(1)]) + \frac{(1-p_f) + (1-p_f)^2(n-2)}{n_0 n^2}(\kappa_{00} - \mathbb{E}[Y_i(0)])\right\} \\
&= b^2\delta p_f(1-p_f)\left(\frac{2}{n_1}\mathbb{E}[Y_i(1)\,|\,f(\boldsymbol{X}_i)=1] + \frac{2}{n_0}\mathbb{E}[Y_i(0)\,|\,f(\boldsymbol{X}_i)=0] + \delta\cdot\frac{n}{n_1 n_0}\right) + O\left(\frac{\delta}{n^2}\right).
\end{aligned}
$$

□

C Difference of the PAPE variances

To compute the difference of the two PAPE variances, we first define the following:

$$A_i = \hat{p}_f Y_i(1) - \hat{p}_f \overline{Y(1)}, \quad B_i = (1-\hat{p}_f) Y_i(0) - (1-\hat{p}_f)\overline{Y(0)},$$
$$C_i = f(\boldsymbol{X}_i) Y_i(1) - \overline{f(\boldsymbol{X}) Y(1)}, \quad D_i = (1-f(\boldsymbol{X}_i)) Y_i(0) - \overline{(1-f(\boldsymbol{X}))Y(0)}.$$

Then, a simple algebraic manipulation yields

$$\mathbb{V}(\hat{\tau}_f(\boldsymbol{Z}_n)) = \frac{n^2}{(n-1)^2}\mathbb{E}\left(\sum_{i=1}^n \frac{A_i^2 + C_i^2 - 2A_iC_i}{n_1(n-1)} + \frac{B_i^2 + D_i^2 - 2B_iD_i}{n_0(n-1)} + \xi\right),$$
$$\mathbb{V}(\hat{\tau}_f^*(\boldsymbol{Z}_n)) = \frac{n^2}{(n-1)^2}\mathbb{E}\left(\sum_{i=1}^n \frac{A_i^2}{n_{r1}(n-1)} + \frac{B_i^2}{n_{r0}(n-1)} + \frac{C_i^2 + D_i^2 + 2C_iD_i}{n_f(n-1)} + \xi\right),$$

where $\xi = \frac{1}{n^2}\{\tau_f^2 - np_f(1-p_f)\tau^2 + 2(n-1)(2p_f-1)\tau_f\tau\}$. Given these expressions, the difference is given by

$$\mathbb{V}(\hat{\tau}_f^*) - \mathbb{V}(\hat{\tau}_f) = \frac{n^2}{(n-1)^2}\mathbb{E}\left(\sum_{i=1}^n \frac{A_i^2(n_1-n_{r1})}{n_{r1}n_1(n-1)} + \frac{B_i^2(n_0-n_{r0})}{n_{r0}n_0(n-1)} + \frac{C_i^2(n_1-n_f)}{n_fn_1(n-1)} + \frac{D_i^2(n_0-n_f)}{n_fn_0(n-1)} + \frac{2C_iD_i}{n_f(n-1)} + \frac{2A_iC_i}{n_1(n-1)} + \frac{2B_iD_i}{n_0(n-1)}\right).$$

Under the assumption that $n_1 = n_0 = n_f = n_r = n/2$ and $n_{r0} = n_{r1} = n/4$, we have

$$\begin{aligned}\mathbb{V}(\hat{\tau}_f^*(\boldsymbol{Z}_n)) - \mathbb{V}(\hat{\tau}_f(\boldsymbol{Z}_n)) &= \frac{2n}{(n-1)^2}\mathbb{E}\left(\sum_{i=1}^n \frac{A_i^2 + B_i^2}{n-1} + \frac{2C_iD_i + 2A_iC_i + 2B_iD_i}{n-1}\right) \\ &= \frac{2n}{(n-1)^2}\Big[\mathbb{E}\{p_f^2S_1^2 + (1-p_f)^2S_0^2\} + 2\mathrm{Cov}(f(\boldsymbol{X}_i)Y_i(1), (1-f(\boldsymbol{X}_i))Y_i(0)) \\ &\quad + 2p_f\mathrm{Cov}(f(\boldsymbol{X}_i)Y_i(1), Y_i(1)) + 2(1-p_f)\mathrm{Cov}((1-f(\boldsymbol{X}_i))Y_i(0), Y_i(0))\Big].\end{aligned}$$

Finally, note the following:

$$\begin{aligned}&\mathrm{Cov}(f(\boldsymbol{X}_i)Y_i(1), (1-f(\boldsymbol{X}_i))Y_i(0)) \\ &= \mathbb{E}\{f(\boldsymbol{X}_i)Y_i(1)(f(\boldsymbol{X}_i)-1)Y_i(0)\} - \mathbb{E}\{f(\boldsymbol{X}_i)Y_i(1)\}\mathbb{E}\{(1-f(\boldsymbol{X}_i))Y_i(0)\} \\ &= -\Pr(f(\boldsymbol{X}_i)=1)\mathbb{E}(Y_i(1)\mid f(\boldsymbol{X}_i)=1)\Pr(f(\boldsymbol{X}_i)=0)\mathbb{E}(Y_i(0)\mid f(\boldsymbol{X}_i)=0) \\ &= -p_f(1-p_f)\mathbb{E}(Y_i(0)\mid f(\boldsymbol{X}_i)=0)\mathbb{E}(Y_i(1)\mid f(\boldsymbol{X}_i)=1),\end{aligned}$$

and

$$p_f\mathrm{Cov}(f(\boldsymbol{X}_i)Y_i(1), Y_i(1)) = p_f^2\{\mathbb{E}(Y_i^2(1)\mid f(\boldsymbol{X}_i)=1) - \mathbb{E}(Y_i(1))\mathbb{E}(Y_i(1)\mid f(\boldsymbol{X}_i)=1)\}$$
$$(1-p_f)\mathrm{Cov}(f(\boldsymbol{X}_i)Y_i(0), Y_i(0)) = (1-p_f)^2\{\mathbb{E}(Y_i^2(0)\mid f(\boldsymbol{X}_i)=0) - \mathbb{E}(Y_i(0))\mathbb{E}(Y_i(0)\mid f(\boldsymbol{X}_i)=0)\}.$$

Hence, we have

$$\begin{aligned}&\mathbb{V}(\hat{\tau}_f^*(\boldsymbol{Z}_n)) - \mathbb{V}(\hat{\tau}_f(\boldsymbol{Z}_n)) \\ &= \frac{2n}{(n-1)^2}\Big[p_f^2\mathbb{V}(Y_i(1)) + (1-p_f)^2\mathbb{V}(Y_i(0)) - 2p_f(1-p_f)\mathbb{E}(Y_i(0)\mid f(\boldsymbol{X}_i)=0)\mathbb{E}(Y_i(1)\mid f(\boldsymbol{X}_i)=1) \\ &\quad + 2p_f^2\{\mathbb{E}(Y_i^2(1)\mid f(\boldsymbol{X}_i)=1) - \mathbb{E}(Y_i(1))\mathbb{E}(Y_i(1)\mid f(\boldsymbol{X}_i)=1)\} \\ &\quad + 2(1-p_f)^2\{\mathbb{E}(Y_i^2(0)\mid f(\boldsymbol{X}_i)=0) - \mathbb{E}(Y_i(0))\mathbb{E}(Y_i(0)\mid f(\boldsymbol{X}_i)=0)\}\Big] \\ &= \frac{2n}{(n-1)^2}[p_f^2\mathbb{V}(Y_i(1)) + (1-p_f)^2\mathbb{V}(Y_i(0)) - 2p_f(1-p_f)\mathbb{E}(Y_i(0)\mid f(\boldsymbol{X}_i)=0)\mathbb{E}(Y_i(1)\mid f(\boldsymbol{X}_i)=1) \\ &\quad + 2p_f^2\{\mathbb{V}(Y_i(1)\mid f(\boldsymbol{X}_i)=1) + (1-p_f)\{\mathbb{E}(Y_i(1)\mid f(\boldsymbol{X}_i)=1) - \mathbb{E}(Y_i(1)\mid f(\boldsymbol{X}_i)=0)\}\mathbb{E}(Y_i(1)\mid f(\boldsymbol{X}_i)=1)\} \\ &\quad + 2(1-p_f)^2\{\mathbb{V}(Y_i(0)\mid f(\boldsymbol{X}_i)=0) + p_f\{\mathbb{E}(Y_i(0)\mid f(\boldsymbol{X}_i)=0) - \mathbb{E}(Y_i(0)\mid f(\boldsymbol{X}_i)=1)\}\mathbb{E}[Y_i(0)\mid f(\boldsymbol{X}_i)=0]\}. \quad \square\end{aligned}$$

D Comparison under the simplifying assumptions

Define $M_{st} = \mathbb{E}(Y_i(s)|f(\boldsymbol{X}_i) = t)$ for $s, t \in \{0, 1\}$. Then, we can rewrite the variance difference as

$$\begin{aligned}\mathbb{V}(\hat{\tau}_f^*(\boldsymbol{Z}_n)) - \mathbb{V}(\hat{\tau}_f(\boldsymbol{Z}_n)) = \frac{2n}{(n-1)^2}\Bigg[& p_f^2\mathbb{V}(Y_i(1)) + (1-p_f)^2\mathbb{V}(Y_i(0)) - 2p_f(1-p_f)M_{11}M_{00} \\ &+ 2p_f^2\{\mathbb{V}(Y_i(1) \mid f(\boldsymbol{X}_i) = 1) + (1-p_f)(M_{11} - M_{10})M_{11}\} \\ &+ 2(1-p_f)^2\{\mathbb{V}(Y_i(0) \mid f(\boldsymbol{X}_i) = 0) + p_f(M_{00} - M_{01})M_{00}\}\Bigg].\end{aligned}$$

Now, consider a constant shift of the outcome variable, i.e., $Y_i(t) + \delta$ for $t = 0, 1$. Then, the variance difference becomes,

$$\begin{aligned}&\mathbb{V}(\hat{\tau}_f^*(\boldsymbol{Z}_n)) - \mathbb{V}(\hat{\tau}_f(\boldsymbol{Z}_n)) \\ &= \frac{2n}{(n-1)^2}\Bigg[p_f^2\mathbb{V}(Y_i(1)) + (1-p_f)^2\mathbb{V}(Y_i(0)) - 2p_f(1-p_f)(M_{11}+\delta)(M_{00}+\delta) \\ &\quad + 2p_f^2\{\mathbb{V}(Y_i(1) \mid f(\boldsymbol{X}_i) = 1) + (1-p_f)(M_{11} - M_{10})(M_{11}+\delta)\} \\ &\quad + 2(1-p_f)^2\{\mathbb{V}(Y_i(0) \mid f(\boldsymbol{X}_i) = 0) + p_f(M_{00} - M_{01})(M_{00}+\delta)\}\Bigg] \\ &= \frac{2n}{(n-1)^2}\Bigg[p_f^2\mathbb{V}(Y_i(1)) + (1-p_f)^2\mathbb{V}(Y_i(0)) - 2p_f(1-p_f)M_{11}M_{00} \\ &\quad + 2p_f^2\{\mathbb{V}(Y_i(1) \mid f(\boldsymbol{X}_i) = 1) + (1-p_f)(M_{11} - M_{10})M_{11}\} \\ &\quad + 2(1-p_f)^2\{\mathbb{V}(Y_i(0) \mid f(\boldsymbol{X}_i) = 0) + p_f(M_{00} - M_{01})M_{00}\} \\ &\quad - 2p_f(1-p_f)\delta^2 + 2p_f(1-p_f)\delta\{p_f(M_{11} - M_{10}) + (1-p_f)(M_{00} - M_{01}) - M_{11} - M_{00}\}\Bigg] \\ &= \frac{2n}{(n-1)^2}\Bigg[p_f^2\mathbb{V}(Y_i(1)) + (1-p_f)^2\mathbb{V}(Y_i(0)) - 2p_f(1-p_f)M_{11}M_{00} \\ &\quad + 2p_f^2\{\mathbb{V}(Y_i(1) \mid f(\boldsymbol{X}_i) = 1) + (1-p_f)(M_{11} - M_{10})M_{11}\} \\ &\quad + 2(1-p_f)^2\{\mathbb{V}(Y_i(0) \mid f(\boldsymbol{X}_i) = 0) + p_f(M_{00} - M_{01})M_{00}\} \\ &\quad - 2p_f(1-p_f)\delta^2 - 2p_f(1-p_f)\delta\{p_f(M_{00} + M_{10}) + (1-p_f)(M_{11} + M_{01})\}\Bigg].\end{aligned}$$

Thus, we observe that the variance difference decreases by

$$2p_f(1-p_f)\delta^2 + 2p_f(1-p_f)\delta\{p_f(M_{00} + M_{10}) + (1-p_f)(M_{11} + M_{01})\}.$$

Since the *ex-ante* estimator is completely unaffected by this change, the constant shift increases the variance of the *ex-post* evaluation estimator by the same amount. Under the simplifying assumptions, we have,

$$M_{11} + M_{01} = M_{00} + M_{10} = 0.$$

Therefore, we can bound the difference in variance from below as follows:

$$\begin{aligned}
\mathbb{V}(\hat{\tau}_f^*(\boldsymbol{Z}_n)) - \mathbb{V}(\hat{\tau}_f(\boldsymbol{Z}_n)) &= \frac{2n}{(n-1)^2}\Big[p_f^2\mathbb{V}(Y_i(1)) + (1-p_f)^2\mathbb{V}(Y_i(0)) - 2p_f(1-p_f)M_{11}M_{00} \\
&\quad + 2p_f^2\{\mathbb{V}(Y_i(1)\mid f(\boldsymbol{X}_i)=1) + (1-p_f)(M_{11}-M_{10})M_{11}\} \\
&\quad + 2(1-p_f)^2\{\mathbb{V}(Y_i(0)\mid f(\boldsymbol{X}_i)=0) + p_f(M_{00}-M_{01})M_{00}\}\Big] \\
&= \frac{2n}{(n-1)^2}\Big[p_f^2\mathbb{V}(Y_i(1)) + (1-p_f)^2\mathbb{V}(Y_i(0)) - 2p_f(1-p_f)M_{11}M_{00} \\
&\quad + 2p_f^2\{\mathbb{V}(Y_i(1)\mid f(\boldsymbol{X}_i)=1) + (1-p_f)(M_{11}+M_{00})M_{11}\} \\
&\quad + 2(1-p_f)^2\{\mathbb{V}(Y_i(0)\mid f(\boldsymbol{X}_i)=0) + p_f(M_{00}+M_{11})M_{00}\}\Big] \\
&= \frac{2n}{(n-1)^2}\Big[p_f^2\mathbb{V}(Y_i(1)) + (1-p_f)^2\mathbb{V}(Y_i(0)) + 2p_f^2\mathbb{V}(Y_i(1)\mid f(\boldsymbol{X}_i)=1) \\
&\quad + 2(1-p_f)^2\mathbb{V}(Y_i(0)\mid f(\boldsymbol{X}_i)=0) + 2p_f(1-p_f)[(1-p_f)M_{00}^2 + p_fM_{11}^2]\Big] \geq 0.
\end{aligned}$$

□

E Outcome model for the numerical study

$$\begin{aligned}
\mathbb{E}(Y_i(t)|\boldsymbol{X}_i) = {} & 1.60 + 0.53 \times x_{29} - 3.80 \times x_{29}(x_{29}-0.98)(x_{29}+0.86) - 0.32 \times \mathbf{1}\{x_{17}>0\} \\
& + 0.21 \times \mathbf{1}\{x_{42}>0\} - 0.63 \times x_{27} + 4.68 \times \mathbf{1}\{x_{27}<-0.61\} - 0.39 \times (x_{27}+0.91)\mathbf{1}\{x_{27}<-0.91\} \\
& + 0.75 \times \mathbf{1}\{x_{30}\leq 0\} - 1.22 \times \mathbf{1}\{x_{54}\leq 0\} + 0.11 \times x_{37}\mathbf{1}\{x_4\leq 0\} - 0.71 \times \mathbf{1}\{x_{17}\leq 0, t=0\} \\
& - 1.82 \times \mathbf{1}\{x_{42}\leq 0, t=1\} + 0.28 \times \mathbf{1}\{x_{30}\leq 0, t=0\} \\
& + \{0.58 \times x_{29} - 9.42 \times x_{29}(x_{29}-0.67)(x_{29}+0.34)\} \times \mathbf{1}\{t=1\} \\
& + (0.44 \times x_{27} - 4.87 \times \mathbf{1}\{x_{27}<-0.80\}) \times \mathbf{1}\{t=0\} - 2.54 \times \mathbf{1}\{t=0, x_{54}\leq 0\}.
\end{aligned}$$