Externally Valid Policy Evaluation from Randomized Trials Using Additional Observational Data

Sofia Ek
Uppsala University
sofia.ek@it.uu.se

Dave Zachariah
Uppsala University
dave.zachariah@it.uu.se

Abstract

Randomized trials are widely considered as the gold standard for evaluating the effects of decision policies. Trial data is, however, drawn from a population which may differ from the intended target population and this raises a problem of external validity (aka. generalizability). In this paper we seek to use trial data to draw valid inferences about the outcome of a policy on the target population. Additional covariate data from the target population is used to model the sampling of individuals in the trial study. We develop a method that yields certifiably valid trial-based policy evaluations under any specified range of model miscalibrations. The method is nonparametric and the validity is assured even with finite samples. The certified policy evaluations are illustrated using both simulated and real data.

1 Introduction

Randomized controlled trials (RCT) are often considered to be the 'gold standard' when evaluating the effects of different decisions or, more generally, decision policies. RCT studies circumvent the need to identify and model potential confounding variables that arise in observational studies. They enable the evaluation and learning of decision policies for use in, e.g., clinical decision support, precision medicine and recommendation systems (Qian and Murphy, 2011; Zhao et al., 2012; Kosorok and Laber, 2019).

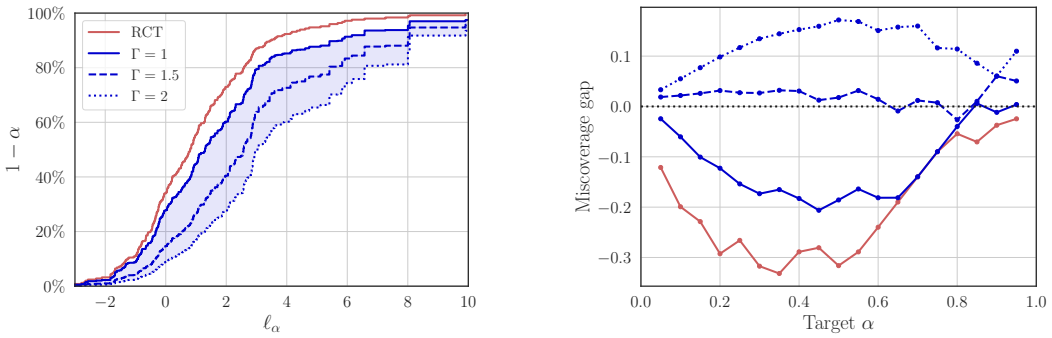
However, RCTs sample individuals that may differ systematically from a target population of interest. For instance, clinical trials usually involve only individuals who do not have any relevant comorbidities and those who volunteer for trials may very well exhibit different characteristics than the target population. Invalid inferences about a decision policy can be potentially harmful in safety-critical applications, where the cautionary principle of "above all, do no harm" applies (Smith, 2005). This is especially challenging since the distributions of population characteristics are unknown. How can we *generalize* results from the trial sample to the intended population?

The focus of this paper is the problem of establishing *externally* valid inferences about outcomes in a target population, when using experimental results from a trial population (Campbell and Stanley, 1963; Manski, 2007; Westreich, 2019). We consider evaluating a decision policy, denoted π, that maps covariates X of an individual onto a recommended action A. The outcome of this decision has an associated loss L (aka. disutility or negative reward). Thus L may directly represent the outcome of interest. We assume the availability of samples (X, A, L) from the trial population and *additional* covariate data X from the target population (Lesko et al., 2016; Li et al., 2022; Colnet et al., 2024). The covariate data is used to model the sampling of individuals in the trial study. We propose a method for evaluating policy π that

- is nonparametric and makes no assumptions about the distributional forms of the data,
- takes into account possible covariate shifts from trial to target distribution, even when using miscalibrated sampling models with unmeasured selection factors,

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2230



(a) (b)

Figure 1: Inferring the out-of-sample losses of a policy π. (a) The loss L is bounded by an upper limit ℓ_α with a probability of at least $1-\alpha$. The RCT-based limit curve uses only trial data, whereas the other limit curves also utilize a sampling model trained using additional covariate data X from the target population. Each limit curve in blue is certified to provide valid inferences for models miscalibrated up to a degree Γ defined in (3). (b) Gap between the actual probability of exceeding the limit, $L > \ell_\alpha$, and the nominal probability of miscoverage α. A negative gap means the inference ℓ_α is *invalid*, while a positive gap implies it is *valid* but conservative. Details of the experiment are presented in Section 5.1.

- and certifies valid finite-sample inferences of the out-of-sample loss, up to any specified degree of model miscalibration.

Many policy evaluation methods are focused on estimating the expected loss $\mathbb{E}_\pi[L]$ of π. However, since a substantial portion of losses L may exceed the mean, this focus can miss important tail events (Wang et al., 2018; Huang et al., 2021). By contrast, evaluating a policy in terms of its out-of-sample loss provides a more complete characterization of its performance and is consonant with the cautionary principle. Figure 1 illustrates the evaluation of π using limit curves which upper bound the out-of-sample loss L with a given probability $1-\alpha$. A limit curve based on RCT-data alone is only ensured to be valid for a trial population. Using additional covariate data, however, we can certify the validity of the inferences for the target population up to any specified degree of miscalibration of the sampling model.

The rest of the paper is outlined as follows. We first state the problem of interest in Section 2 and relate it to the existing literature in Section 3. We then propose a policy evaluation method in Section 4 and demonstrate its properties using both synthetic and real data in Section 5. We conclude the paper with a discussion about the properties of the method in Section 6 and its broader impact in Section 7.

2 Problem Formulation

Any policy π, whether deterministic or randomized, can be described by a distribution $p_\pi(A|X)$. Each covariate X, unmeasured selection factor U and action A has an associated loss $L \in (-\infty, L_{\max})$. We consider here a discrete action space, i.e. $A \in \{1 \dots, K\}$. The decision process has a causal structure that can be formalized by a directed acyclic graph, visually summarized in Figure 2 (Peters et al., 2017). The sampling indicator S indicates whether individuals are drawn from a *target* population, $S = 0$, or a *trial* population, $S = 1$. The causal structure allows us to decompose the two distributions. Specifically, the target distribution factorizes as

$$p_\pi(X, U, A, L|S = 0) = p(X, U|S = 0) \cdot p_\pi(A|X) \cdot p(L|X, U, A), \tag{1}$$

where only the policy $p_\pi(A|X)$ is known. Similarly, the trial distribution factorizes as

$$p(X, U, A, L|S = 1) = p(X, U|S = 1) \cdot p(A|X) \cdot p(L|X, U, A), \tag{2}$$

where only the randomization mechanism $p(A|X)$ is known. In general the characteristics of target and trial populations may *differ*, that is, $p(X, U|S = 0) \neq p(X, U|S = 1)$. The unmeasured U may include self-selection factors that are challenging to record. Note, however, that only factors that also affect the loss L are relevant here.

(a) (b)

Figure 2: Causal structure of process (a) under policy π as well as (b) the trial study. Sampling indicator S distinguishes between the two. For the important case of RCT, assignment of A is not influenced by any covariates so that the path $X \to A$ is eliminated.

From the trial distribution (2), we sample m individuals $\mathcal{D} = \big((X_i, A_i, L_i)\big)_{i=1}^{m}$ independently. In addition, we also obtain n independent samples of *covariate-only* data $\big(X_1, X_2, \ldots, X_n\big)$ from the target population (1). Our aim is to infer the out-of-sample loss L_{n+1} for individual $n+1$ under any policy π. Specifically, we seek a loss limit ℓ_α as a function of $1-\alpha$, such that $L_{n+1} \leq \ell_\alpha$ holds with probability $1-\alpha$ as illustrated by the 'limit curves' in Figure 1a.

The sampling pattern of individuals is described by $p(S|X,U)$. This distribution is unknown, but we assume that a model $\widehat{p}(S|X)$ is available. This model was fitted using held-out data $\{(X_j, S_j)\}$ employing either the conventional logistic model or any state-of-the-art machine learning models (as exemplified below). It can also be obtained from previous studies. There is, however, no guarantee that $\widehat{p}(S|X)$ is calibrated and it may indeed diverge from the unknown sampling pattern. Nevertheless, we want inferences about the out-of-sample loss to be valid also for miscalibrated models. We therefore express the degree of miscalibration in terms of the selection odds:

$$\frac{1}{\Gamma} \;\leq\; \underbrace{\frac{p(S=0|X,U)}{p(S=1|X,U)}}_{\text{unknown selection odds}} \Big/ \underbrace{\frac{\widehat{p}(S=0|X)}{\widehat{p}(S=1|X)}}_{\text{nominal selection odds}} \;\leq\; \Gamma \qquad (a.s.) \tag{3}$$

That is, the nominal selection odds can diverge by a factor Γ, where $\Gamma = 1$ implies a perfectly calibrated model. This model includes all sources of errors (selection bias, model misspecification, estimation error).

A limit ℓ_α^Γ provides an *externally valid* inference of L_{n+1}, up to any specified degree of miscalibration Γ, if it satisfies

$$\boxed{\mathbb{P}_\pi\Big\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}) \mid S=0\Big\} \geq 1-\alpha, \quad \forall\alpha.} \tag{4}$$

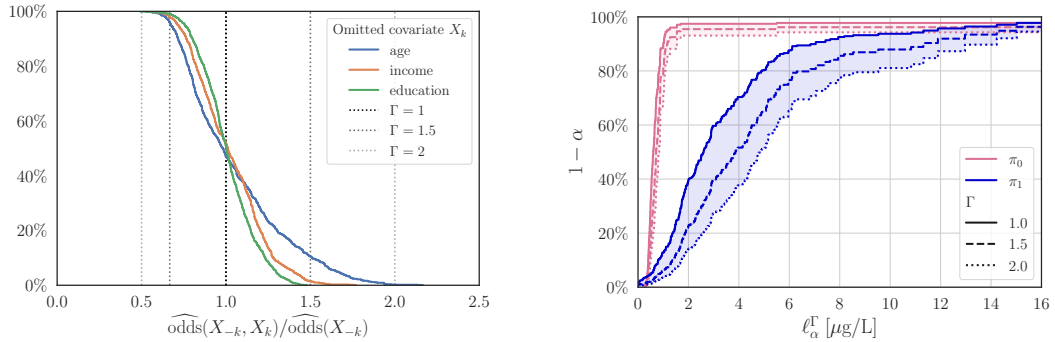
The problem we consider is to construct this externally valid limit ℓ_α^Γ. This limit allows us to infer the full loss distribution of a future individual with L_{n+1} under policy π, rather than merely the expected loss $\mathbb{E}_\pi[L]$. Specifically, the tail losses are important in healthcare and other safety critical applications where erroneous inferences could be harmful, and a cautious approach when implementing new policies is needed.

The limit curve ℓ_α^Γ for policy π is valid up to any declared degree of odds miscalibration Γ, which establishes the credibility of the policy evaluation, cf. Manski (2003). While Γ is unknowable, especially under unmeasured U, we can use ideas from sensitivity analysis to guide its selection using measured data (Rosenbaum and Rubin, 1983; Tan, 2006) . Following the method in Huang (2024), we treat observed selection factors in X as unmeasured U in (3) to benchmark appropriate values for Γ, as detailed in subsection 4.1.

By increasing the range of Γ, we certify the validity of the inference under increasingly credible assumptions on $\widehat{p}(S|X)$ (Manski, 2003). As the model credibility increases, however, the informativeness of the inferences decreases. Since the upper bound on the losses, $L_{\max}$, is a trivial and uninformative limit, we may define the informativeness of ℓ_α^Γ as

$$\text{Informativeness} = 1 - \inf\{\alpha : \ell_\alpha^\Gamma(\mathcal{D}) < L_{\max}\}. \tag{5}$$

https://doi.org/10.52202/079017-2230



(a) (b)

Figure 3: (a) Omitting measured selection factors to benchmark credible values for Γ in (3). (b) Inferred blood mercury levels [μg/L] in a target population under 'high' and 'low' seafood consumption (π_1 and π_0, respectively). Limit curves for degrees of odds miscalibration $\Gamma \in [1, 2]$.

That is, the right limit of a limit curve, which decreases with the degree of miscalibration. Figure 1a shows curves that are valid for miscalibration in the range $\Gamma \in [1, 2]$, where the informativeness is 95% and 90%, respectively. The latter figure means that we can infer a non-trivial bound on the loss for 90% of the target population.

Remark: This paper addresses the evaluation of a given policy, whether proposed by a clinical expert or learned from historical data. By setting aside samples from an RCT study, a policy π can be learned, and its out-of-sample performance evaluated using the proposed methodology.

3 Background Literature

The problem considered herein is related to the problem of causal inference when combining data from randomized controlled trials and observational studies. Examples of the setting where only covariate data from the observational study is available can be found in Lesko et al. (2016) and Li et al. (2022). For additional examples, refer to the survey by Colnet et al. (2024). Within the broader area of generalizability and transportability, the problem represents the case were the sampling probability depends solely on the covariates X, and is independent of the action and the loss (Pearl and Bareinboim, 2014; Lesko et al., 2017; Degtiar and Rose, 2023). The problem is also related to a broader literature of statistical learning under covariate shifts, see for instance Shimodaira (2000); Sugiyama et al. (2007); Quinonero-Candela et al. (2008); Reddi et al. (2015); Chen et al. (2016).

The most common object of inference in policy evaluation is the expected loss $\mathbb{E}_\pi[L]$ and a popular method for estimating it is inverse probability of sampling weighting (IPSW), which models covariate shifts from trial to target populations. The estimator using RCT-data is defined as

$$V_{\text{IPSW}}^\pi = \frac{1}{n} \sum_{i=1}^{m} \frac{\widehat{p}(S=0|X_i)}{\widehat{p}(S=1|X_i)} \cdot \frac{p_\pi(A_i|X_i)}{p(A_i)} \cdot L_i. \tag{6}$$

This methodology has been applied in various studies, see for example Cole and Stuart (2010); Stuart et al. (2011); Westreich et al. (2017); Buchanan et al. (2018). It is widely recognized that misspecified logistic models can introduce bias when estimating the weights (Colnet et al., 2024) and more recent works have suggested using flexible models, such as generalized boosted methods (Kern et al., 2016), instead. The counterpart to IPSW, used in off-policy evaluation with observational data, is inverse propensity weighting (IPW). The problem with misspecified logistic models also applies here when estimating the classification probabilities, aka. propensity scores (McCaffrey et al., 2004; Lee et al., 2010). In this case, generalized boosted methods and covariate-balancing methods have shown to be promising alternatives (Setodji et al., 2017; Tu, 2019). The literature on IPSW and IPW is mainly focused on average treatment effect estimation, that is $\mathbb{E}_{\pi_1}[L] - \mathbb{E}_{\pi_0}[L]$ where π_1 and π_0 denote the 'treat all' and 'treat none' policies, respectively. By contrast, we want to certify the distributional properties of L_{n+1} for any π, even under miscalibration of $\widehat{p}(S|X)$.

Conformal prediction is a distribution-free methodology focused on creating covariate-specific prediction regions that are valid for finite-samples (Vovk et al., 2005; Shafer and Vovk, 2008). The methodology was extended by Tibshirani et al. (2019) to also work for known covariate shifts. Jin et al. (2023) combined the marginal sensitivity methodology developed in Tan (2006) with the conformal prediction for covariate shifts to perform sensitivity analysis of treatment effects in the case of unobserved confounding. Our methodology draws upon techniques in conformal prediction, but instead of providing covariate-specific prediction intervals under a policy π, we are concerned with evaluating *any* π over a target population.

When full identification of the causal effect is not possible due to unmeasured confounders, partial identification sensitivity analysis can be used to evaluate the robustness of the estimates. Rosenbaum and Rubin (1983) introduced a sensitivity parameter to bound the odds ratio of the probability of treatment. More recent work has extended this approach to account for treatment effect heterogeneity and non-binary treatments, as seen in Tan (2006); Shen et al. (2011); Zhao et al. (2019); Dorn and Guo (2023) among others. However, it is often challenging to interpret the absolute value of the sensitivity parameter in practical scenarios. To address this, recent research has proposed benchmarking, or calibrating, results by estimating the effects of unmeasured confounders, see for example Hsu and Small (2013); Franks et al. (2019); Veitch and Zaveri (2020); De Bartolomeis et al. (2024). Note that all these papers work with sensitivity analysis for observational studies. Huang (2024) instead combines sensitivity analysis for generalization with benchmarking to determine reasonable values of Γ.

A biased sample selection, similar to that described by (3), was considered in the context of average treatment effect estimation by Nie et al. (2021). In contrast to that work, our primary focus is on ensuring the validity of inferences regarding out-of-sample losses, even when dealing with finite training data. We achieve this using a sample-splitting technique.

4 Method

Here we construct a limit $\ell_\alpha^\Gamma(\mathcal{D})$ on the out-of-sample losses under policy π that satisfies (4) for any given specified degree of miscalibration Γ.

4.1 Benchmarking degree of miscalibration

The limit curve $\ell_\alpha^\Gamma(\mathcal{D})$ holds up to the specified degree of odds miscalibration Γ. Although Γ is inherently unknown, sensitivity analysis and methods for assessing the calibration of models can guide its estimation using available data.

We start with a benchmarking method that specifically accounts for the potential impact of unmeasured selection factors, U. Building on the approach in Huang (2024), we treat observed selection factors in X as proxies for unmeasured U in equation (3), providing a benchmark for selecting suitable Γ values. Specifically, let the the omitted selection factor X_k be U and X_{-k} denotes the remaining covariates. Then the ratio in (3) is estimated by dividing $\widehat{\text{odds}}(X_{-k}, X_k) = \frac{\widehat{p}(S=0|X)}{\widehat{p}(S=1|X)}$ by $\widehat{\text{odds}}(X_{-k}) = \frac{\widehat{p}(S=0|X_{-k})}{\widehat{p}(S=1|X_{-k})}$. Figure 3a summarizes the distribution of $\widehat{\text{odds}}(X_{-k}, X_k)/\widehat{\text{odds}}(X_{-k})$ using a real data set, where the omitted X_k is either age, income or education. We see that the corresponding ratios all fall within $\Gamma = 2$. We can therefore conclude that if any unmeasured selection factor U is weaker than the omitted factors, it is credible to set Γ in the range of 1.5 to 2. The corresponding loss curves are shown in Figure 3b, which illustrates the blood mercury levels in a target population under policies of 'high' respective 'low' seafood consumption that can credibly be extrapolated from trial data. More details are available in Section 5.2. Note that this benchmarking method serves as a guide for assessing the influence of unmeasured selection factors U (Cinelli and Hazlett, 2019).

Without unmeasured selection factors, methods for assessing the calibration of models $\widehat{p}(S|X)$ discussed in, e.g., Murphy and Winkler (1977); Naeini et al. (2015); Widmann et al. (2019) can guide the specified lower bound of the range of miscalibration. The nominal selection odds in (3), i.e., $\widehat{\text{odds}}(X) = \frac{\widehat{p}(S=0|X)}{\widehat{p}(S=1|X)}$, can be quantized into several bins and for each bin the unknown selection odds, i.e., $\text{odds}(X) = \frac{p(S=0|X)}{p(S=1|X)}$, can be estimated by counting the samples from both the target and trial distributions present. In the case of a well-calibrated model, estimated unknown odds should

 https://doi.org/10.52202/079017-2230

match the quantized nominal odds for each bin. To assess calibration, this process can be iterated across multiple ranges within the dataset and visualized through a reliability diagram, as exemplified in Figure 4. Using (3), we have that

$$\frac{1}{\Gamma} \cdot \widehat{\text{odds}}(X) \;\leq\; \text{odds}(X) \;\leq\; \Gamma \cdot \widehat{\text{odds}}(X).$$

Take the expectation with respect to X, conditional on the nominal odds in a specified interval (or bin) I, so that:

$$\frac{1}{\Gamma} \cdot \mathbb{E}[\widehat{\text{odds}}(X) \mid \widehat{\text{odds}}(X) \in I] \;\leq\; \mathbb{E}[\text{odds}(X) \mid \widehat{\text{odds}}(X) \in I] \;\leq\; \Gamma \cdot \mathbb{E}[\widehat{\text{odds}}(X) \mid \widehat{\text{odds}}(X) \in I].$$

The expected odds is then estimated for each bin I by counting samples from the target and trial distributions.

4.2 Inferring out-of-sample limit

We will now construct the limit $\ell_\alpha^\Gamma(\mathcal{D})$. For this we need to describe the distribution shift from trial to target distribution for all samples, including the $n+1$ sample. We begin by considering the true distribution shift expressed using the ratio

$$\frac{p_\pi(X, U, A, L|S=0)}{p(X, U, A, L|S=1)}. \tag{7}$$

Inserting the factorizations (1) and (2) into this ratio shows that specifying the distribution shift requires the unknown (conditional) covariate distribution $p(X, U|S)$. We can, however, bound (7) using the model of the sampling pattern, $\widehat{p}(S|X)$, as follows:

$$c \cdot \underline{W}^\Gamma \leq \frac{p_\pi(X, U, A, L|S=0)}{p(X, U, A, L|S=1)} \leq c \cdot \overline{W}^\Gamma, \tag{8}$$

where

$$\underline{W}^\Gamma = \frac{1}{\Gamma} \cdot \frac{\widehat{p}(S=0|X)}{\widehat{p}(S=1|X)} \cdot \frac{p_\pi(A|X)}{p(A|X)}, \qquad \overline{W}^\Gamma = \Gamma \cdot \frac{\widehat{p}(S=0|X)}{\widehat{p}(S=1|X)} \cdot \frac{p_\pi(A|X)}{p(A|X)}, \tag{9}$$

and $c = \frac{p(S=1)}{p(S=0)}$ is a constant. To see this, we note that the ratio in (7) can be expressed as

$$c \cdot \frac{p(S=0|X, U)}{p(S=1|X, U)} \cdot \frac{p_\pi(A|X)}{p(A|X)},$$

using Bayes' rule. The bound (8) follows by applying (3). We proceed to show that the factors (9) are sufficient to construct an externally valid limit ℓ_α^Γ for odds divergences up to degree Γ, similar to Ek et al. (2023).

To ensure finite-sample guarantees, the trial data is randomly divided into two sets, $\mathcal{D} = \mathcal{D}' \cup \mathcal{D}''$, with respective samples sizes of m' and $m - m'$. The set $\mathcal{D}'$ is used to construct

$$\overline{w}_\beta^\Gamma(\mathcal{D}') = \begin{cases} \overline{W}_{[\lceil (m'+1)(1-\beta) \rceil]}^\Gamma, & (m'+1)(1-\beta) \leq m', \\ \infty, & \text{otherwise}, \end{cases} \tag{10}$$

where $\overline{W}_{[\cdot]}^\Gamma$ is the upper limit in (9) evaluated over $\mathcal{D}'$ and ordered $\overline{W}_{[1]}^\Gamma \leq \overline{W}_{[2]}^\Gamma \leq \cdots \leq \overline{W}_{[m']}^\Gamma$. We show that (10) upper bounds the ratio (7) for a future sample with probability $1 - \beta$ for any choice of $\beta \in (0, 1)$. The set $\mathcal{D}''$ is used to construct a stand-in for the unknown cumulative distribution function of the out-of-sample loss:

$$\widehat{F}(\ell; \mathcal{D}'', w) = \frac{\sum_{i \in \mathcal{D}''} \underline{W}_i^\Gamma \mathbb{1}(L_i \leq \ell)}{\sum_{i \in \mathcal{D}''} \left[\underline{W}_i^\Gamma \mathbb{1}(L_i \leq \ell) + \overline{W}_i^\Gamma \mathbb{1}(L_i > \ell) \right] + w}, \tag{11}$$

where $w > 0$ is a free variable representing the unknown out-of-sample weight $\overline{W}_{n+1}^\Gamma$ for a future sample. Based on (10) and (11), define $\ell_{\alpha,\beta}^\Gamma$ as the quantile function

$$\ell_{\alpha,\beta}^\Gamma = \inf \left\{ \ell : \widehat{F}(\ell; \mathcal{D}'', \overline{w}_\beta^\Gamma(\mathcal{D}')) \geq \frac{1-\alpha}{1-\beta} \right\}. \tag{12}$$

This enables us to construct a valid limit on the future loss L_{n+1} for any miscoverage probability $\alpha \in (0, 1)$.

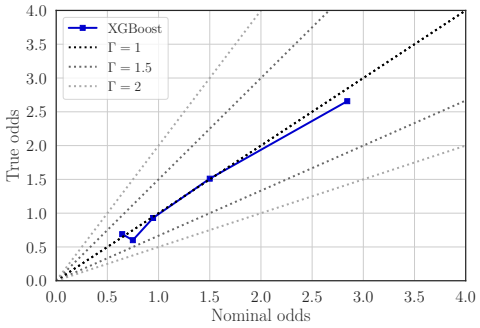


Figure 4: Reliability diagram of the observed odds against the average predicted nominal odds obtained from models $\widehat{p}(S|X)$.

Algorithm 1 A set of limit curves for policy π

input Model $\widehat{p}(S|X)$, policy $p_\pi(A|X)$, trial policy $p(A|X)$, a set of miscalibration degrees $\{1, \ldots, \Gamma_{\text{max}}\}$ and trial data $\mathcal{D}$.
output $\{(\Gamma, \alpha, \ell_\alpha^\Gamma)\}$

Randomly split $\mathcal{D}$ into $\mathcal{D}'$ and $\mathcal{D}''$.
for $\Gamma \in \{1, \ldots, \Gamma_{\text{max}}\}$ **do**
  for $\alpha \in \{0, \ldots, 1\}$ **do**
    for $\beta \in \{0, \ldots, \alpha\}$ **do**
      Compute $\overline{w}_\beta^\Gamma$ as in (10).
      Compute $\ell_{\alpha,\beta}^\Gamma$ as in (12).
    end for
    Set ℓ_α^Γ to the smallest $\ell_{\alpha,\beta}^\Gamma$.
    Save $(\Gamma, \alpha, \ell_\alpha^\Gamma)$.
  end for
end for

Theorem 4.1. *For any odds miscalibration up to degree* Γ*,*

$$\ell_\alpha^\Gamma(\mathcal{D}) = \min_{\beta:0<\beta<\alpha} \ell_{\alpha,\beta}^\Gamma, \tag{13}$$

is an externally valid limit on the out-of-sample loss L_{n+1} *of policy* π*. That is,* (13) *is certified to satisfy* (4)*.*

The method seeks the level β for the bound (10) that yields the tightest limit. The proof is presented in the supplementary material and builds on several techniques developed in Vovk et al. (2005); Tibshirani et al. (2019); Jin et al. (2023); Ek et al. (2023).

Algorithm 1 summarizes the implementation of a set of limit curves given a model of the sampling pattern $\widehat{p}(S|X)$ and a set of miscalibration degrees $[1, \Gamma_{\text{max}}]$. Note that (10) and (11) are step functions in β respective ℓ. Therefore (12) and (13) can be solved by computing the functions at a grid of points. Calculating (10) and (11) requires the sorting of weights, but the sorting operation is a one-time requirement.

Increasing the degree of miscalibration Γ results in a decrease in $\underline{W}^\Gamma$ and an increase in $\overline{W}^\Gamma$ in (9). As weights associated with lower and higher losses decrease and increase, respectively, in (11), the resulting limit $\ell_\alpha^\Gamma(\mathcal{D})$ becomes more conservative.

Remark 1: If the trial population is a small subgroup of the target population (for example in the case of weak overlap), the nominal odds $\frac{\widehat{p}(S=0|X)}{\widehat{p}(S=1|X)}$ will tend to be high. As this yields a significant weight $\overline{w}_\beta^\Gamma(\mathcal{D}')$, the informativeness (5) of $\ell_\alpha^\Gamma(\mathcal{D})$ diminishes. Nevertheless, the guarantee in (4) holds.

Remark 2: The split of $\mathcal{D}$ can be performed in a sample-efficient manner in the case of RCT, where actions are randomized so that $p(A|X) \equiv p(A)$: For the ith sample, (X_i, A_i, L_i), draw $\widetilde{A}_i \sim p_\pi(A|X_i)$. Include sample i in $\mathcal{D}''$ if $A_i = \widetilde{A}_i$, otherwise include it in $\mathcal{D}'$. This sample splitting method ensures that inferences on the loss are based on actions that match those of the policy.

5 Numerical Experiments

We will use both synthetic and real-world data to illustrate the main concepts of policy evaluation with limit curves $(\alpha, \ell_\alpha^\Gamma)$. As a performance benchmark, we estimate the quantile – which yields the tightest limit – using the inverse probability of sampling weighting (Colnet et al., 2024)

$$\ell_\alpha(\mathcal{D}) = \inf\left\{\ell : \widehat{F}_{\text{IPSW}}(\ell; \mathcal{D}) \geq 1 - \alpha\right\}, \tag{14}$$

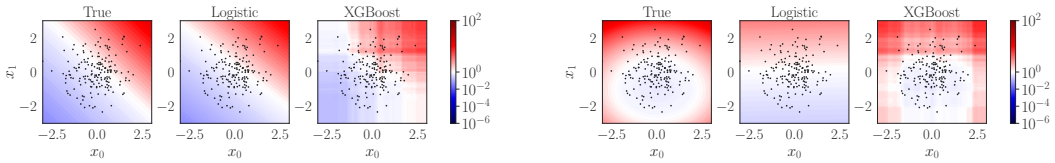
where

$$\widehat{F}_{\text{IPSW}}(\ell; \mathcal{D}) = \frac{1}{m} \sum_{i=1}^{m} \frac{m}{n} \cdot \frac{\widehat{p}(S=0|X_i)}{\widehat{p}(S=1|X_i)} \cdot \frac{p_\pi(A_i|X_i)}{p(A_i)} \cdot \mathbb{1}(L_i \leq \ell).$$

https://doi.org/10.52202/079017-2230

Table 1: Means and variances of covariate distribution $p(X, U|S)$ in (15).

Population	$\mu_{0,S}$	$\mu_{1,S}$	$\mu_{U,S}$	$\sigma^2_{0,S}$	$\sigma^2_{1,S}$	$\sigma^2_{U,S}$
`A` $(S=0)$	0.5	0.5	0.5	1.0	1.0	1.0
`B` $(S=0)$	0.0	0.5	0.0	1.25	1.5	1.25
`Trial` $(S=1)$	0.0	0.0	0.0	1.0	1.0	1.0



(a) Selection odds for target population `A`.

(b) Selection odds for for target population `B`.

Figure 5: Odds $p(S=0|X)/p(S=1|X)$ compared with nominal odds obtained from logistic and XGBoost models $\widehat{p}(S|X)$. The dots are a random subsample of the trial samples.

This is similar to the approach in Huang et al. (2021) but adapted to problems involving data from trial and observational studies.

We examine the impact of increasing the credibility of our model assumptions, i.e., by increasing the miscalibration degree Γ, on the informativeness (5) of the limit curve. In addition, for the simulated data, we also assess the miscoverage gap of the curves

$$\text{Miscoverage gap} = \alpha - \mathbb{P}_\pi\{L_{n+1} > \ell_\alpha(\mathcal{D})|S{=}0\},$$

where a negative gap indicates an invalid limit.

5.1 Illustrations using synthetic data

We consider target and trial populations of individuals with two-dimensional covariates, distributed as follows:

$$X|S = \begin{bmatrix} X_{0,S} \\ X_{1,S} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{0,S} \\ \mu_{1,S} \end{bmatrix}, \begin{bmatrix} \sigma^2_{0,S} & 0 \\ 0 & \sigma^2_{1,S} \end{bmatrix} \right), \qquad U|S \sim \mathcal{N}(\mu_{U,S}, \sigma^2_{U,S}), \tag{15}$$

where the parameters are given in Table 1. The distributions for populations `A`, `B` and `Trial` are taken to be unknown.

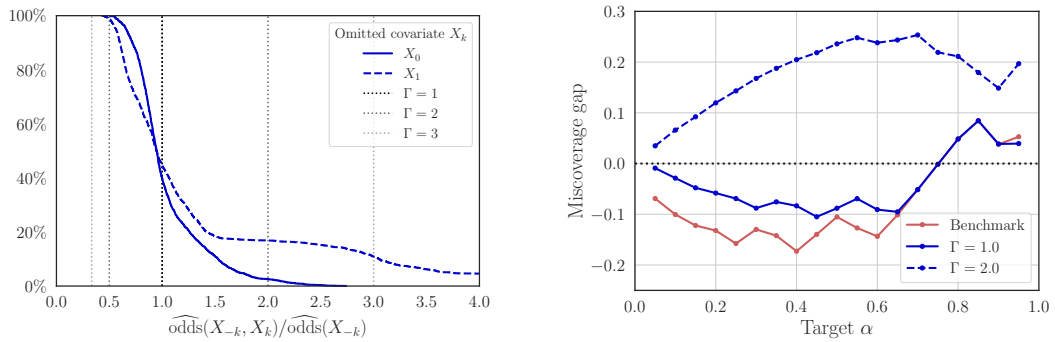
The actions are binary $A \in \{0, 1\}$ and corresponds to 'do not treat' versus 'treat'. We evaluate the 'treat all' policy, i.e. $p_{\pi_1}(A = 1|X) = 1$. The trial is an RCT with equal probability of assignment, i.e., $p(A) \equiv 0.5$. The unknown conditional loss distribution is given by

$$L|(A, X, U) \sim \mathcal{N}(A \cdot X_0^2 + X_1 + A \cdot U + (1 - A), 1).$$

The sampling probability $p(S|X, U)$ is treated as unknown. The complete set of hyperparameters used is provided in the supplementary material.

For the observational data $\big(X_i\big)_{i=1}^n$, we drew $n = 2000$ samples. For the trial data we drew 1000 samples: $m = 500$ samples and $\big((X_i, A_i, L_i)\big)_{i=1}^m$ was used to compute the limit curves. The remaining 500 samples were used to train $\widehat{p}(S|X)$.

Figure 5a compares nominal selection odds obtained from the fitted models with the unknown odds $p(S=0|X)/p(S=1|X)$ for target population `A`. We consider two different fitted models $\widehat{p}(S|X)$: a logistic model, which is conventionally used in the causal inference literature (Westreich et al., 2017), and the more flexible tree-based ensemble model trained by XGBoost (Chen and Guestrin, 2016). In this case, the logistic model happens to be well-specified so that the learned odds approximate the true ones well. The more flexible XGBoost model also provide visually similar odds, albeit less accurate. By contrast, Figure 5b repeats the same exercise for target population `B`. Here the logistic model is misspecified and severely miscalibrated while the XGBoost model continues to provide



(a) (b)

Figure 6: Evaluating a 'treat all' policy π_1 for target population B. (a) Benchmarking the degree of miscalibration Γ using omitted covariates. (b) Miscoverage gaps when degree of miscalibration $\Gamma \in [1, 2]$.

visually similar odds. A well-performing and flexible model is required for a meaningful benchmark of the upper value of Γ. Therefore, we will use the XGBoost model.

In Figure 4, we use the reliability diagram technique to assess the performance of the XGBoost model $\widehat{\text{odds}}(X)$ of the nominal odds for target population B. The model is close to the diagonal suggesting that it is sufficiently flexible to accurately model the odds. This is in line with the result in Figure 5b. We then use observed covariates to calibrate appropriate upper bounds for Γ. Figure 6a shows the evaluation with respect to population B. Assuming that the unmeasured U have no greater selection effect than X_0, a degree of miscalibration $\Gamma = 2$ is a credible choice as it covers more than 95% of the ratios. Since the data is simulated, we evaluate the miscoverage gap of the limit curves of the 'treat all' policy π_1 in Figure 6b for the benchmark and the limit curves for the proposed method. The gap is estimated using 1000 independent runs and for each run drawing 1000 independent new samples $(X_{n+1}, U_{n+1}, A_{n+1}, L_{n+1})$. We see that the benchmark and the limit curve for $\Gamma = 1$ yields a negative miscoverage gap. As the degree of miscalibration Γ increases to 2, the limit curves exhibit positive miscoverage gaps.

The evaluation of π_1 with respect to population A is shown in Figure 1. Results for the logistic model, another policy π_0, and for additional populations are available in the supplementary material.

5.2 Evaluating seafood consumption policies

To illustrate the application of policy evaluation with real data, we study the impact of seafood consumption on blood mercury levels with data from the 2013-2014 National Health and Nutrition Examination Survey (NHANES). Following Zhao et al. (2019), each individual's data includes eight covariates, encompassing gender, age, income, the presence or absence of income information, race, education, smoking history, and the number of cigarettes smoked in the last month. All covariates, except for smoking history U, are treated as measured and denoted X. We excluded one individual due to missing education data and seven individuals with incomplete smoking data. We impute the median income for 175 individuals with no income information. The continuous covariates – age, income and number of cigarettes smoked in the last month – are standardized. After the preprocessing, our data set comprises 1107 individuals. The data is then split into observational data $\mathcal{D}_0$ and trial data $\mathcal{D}_1$ based on the covariates gender, age, income and smoking history (more details are available in the supplementary material) resulting in 646 samples in the observational data and 461 samples in the trial data. The action A describes individual fish or shellfish consumption, categorizing an individual as having either low ($\leq$1 serving in the past month) or high ($>$12 servings in the past month) consumption. The loss L represents the total concentration of blood mercury (measured in μg/L).

To generate counterfactual actions and losses, we consider a balanced RCT so that $p(A) \equiv 0.5$ and learn a model of $p(L|A, X, U)$ from the data using gradient boosting (Freund and Schapire, 1997;

https://doi.org/10.52202/079017-2230

Friedman et al., 2000). Thus during training, the trial data consists of samples (X_i, A_i, L_i) whereas the observational data only contains X.

In Figure 3a we use observed covariates to benchmark appropriate upper bounds for Γ. We exclude each of the seven covariates one by one, highlighting the three most dominant ones in the figure. If the unmeasured covariates have weaker influence than these, a Γ value in the range of 1.5 to 2 is appropriate. In Figure 3b we compare the limit curve for a policy π_0 corresponding to low $(A \equiv 0)$ seafood consumption with the limit curve for a policy π_1 corresponding to high $(A \equiv 1)$ seafood consumption. We use an XGBoost-trained model $\widehat{p}(S|X)$. For reference, a mercury level of 8 μg/L is guidance limit for women of child-bearing age. We see that under a low consumption policy a lower mercury level can be certified for miscalibrated odds $\Gamma \in [1, 2]$. In this case, we can infer a 90% probability that the out-of-sample mercury level falls below the reference value of 8 μg/L.

6 Discussion

We have proposed a method for establishing externally valid policy evaluation based on experimental results from trial studies. The method is nonparametric, making no assumptions on the distributional forms of the data. Using additional covariate data from a target population, it takes into account possible covariate shifts between target and trial populations, and certifies valid finite-sample inferences of the out-of-sample loss L_{n+1}, up to any specified degree of model miscalibration.

Conventional policy evaluation methods focus on $\mathbb{E}_\pi[L]$ and can easily introduce a bias without the user's awareness, particularly when the model of the sampling pattern $\widehat{p}(S|X)$ is misspecified. Lacking any control for miscalibration undermines the possibility to establish external validity. In safety-critical applications, making invalid inferences about a decision policy can be potentially harmful. Hence, adhering to the cautionary principle of "above all, do no harm" is important. The proposed method is designed with this principle in mind, and the limit curve represents the worst-case scenario for the selected degree of miscalibration Γ.

We also exemplify how the reliability diagram technique and the benchmark approach of omitting observed selection factors facilitate a more systematic guidance on the specification of the odds miscalibration degree Γ in any given problem.

7 Broader Impact and Limitations

The method we propose is designed for safety-critical applications, such as clinical decision support, with the cautionary principle in mind. We believe that it offers a valuable tool for policy evaluation in such scenarios. Our approach focuses on limit curves, coupled with a statistical guarantee, for a more detailed understanding of the out-of-sample loss. This facilitates a fairer evaluation by bringing attention to sensitive covariates in the tails. However, it remains important to be aware of biases, and it might be necessary to address them separately to prevent their replication. It is also important to note that the method requires independent samples, a condition that may not be met during major virus outbreaks or similar situations.

Acknowledgments and Disclosure of Funding

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the Swedish Research Council under contract 2021-05022.

References

Ashley L Buchanan, Michael G Hudgens, Stephen R Cole, Katie R Mollan, Paul E Sax, Eric S Daar, Adaora A Adimora, Joseph J Eron, and Michael J Mugavero. Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(4):1193–1209, 2018.

Donald T Campbell and Julian C Stanley. *Experimental and Quasi-experimental Designs for Research*. R. McNally College Publishing Company, 1963.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Xiangli Chen, Mathew Monfort, Anqi Liu, and Brian D Ziebart. Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pages 1270–1279. PMLR, 2016.

Carlos Cinelli and Chad Hazlett. Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67, 12 2019. ISSN 1369-7412. doi: 10.1111/rssb.12348. URL `https://doi.org/10.1111/rssb.12348`.

Stephen R Cole and Elizabeth A Stuart. Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. *American journal of epidemiology*, 172(1):107–115, 2010.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science*, 39(1):165–191, 2024.

Piersilvio De Bartolomeis, Javier Abad Martinez, Konstantin Donhauser, and Fanny Yang. Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. In *International Conference on Artificial Intelligence and Statistics*, pages 1045–1053. PMLR, 2024.

Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10:501–524, 2023.

Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544):2645–2657, 2023.

Sofia Ek, Dave Zachariah, Fredrik D Johansson, and Peter Stoica. Off-policy evaluation with out-of-sample guarantees. *Transactions on Machine Learning Research*, 2023.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2013. URL `https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013`.

AlexanderM Franks, Alexander D'Amour, and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 2019.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2): 337–407, 2000.

Jesse Y Hsu and Dylan S Small. Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics*, 69(4):803–811, 2013.

Audrey Huang, Liu Leqi, Zachary Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment in contextual bandits. *Advances in Neural Information Processing Systems*, 34:23714–23726, 2021.

Melody Y Huang. Sensitivity analysis for the generalization of experimental results. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page qnae012, 03 2024.

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6), 2023.

Holger L Kern, Elizabeth A Stuart, Jennifer Hill, and Donald P Green. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of research on educational effectiveness*, 9(1):103–127, 2016.

 https://doi.org/10.52202/079017-2230

Michael R Kosorok and Eric B Laber. Precision medicine. *Annual review of statistics and its application*, 6:263–286, 2019.

Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, 2010.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Catherine R Lesko, Stephen R Cole, H Irene Hall, Daniel Westreich, William C Miller, Joseph J Eron, Jianmin Li, Michael J Mugavero, and CNICS Investigators. The effect of antiretroviral therapy on all-cause mortality, generalized to persons diagnosed with hiv in the usa, 2009–11. *International journal of epidemiology*, 45(1):140–150, 2016.

Catherine R Lesko, Ashley L Buchanan, Daniel Westreich, Jessie K Edwards, Michael G Hudgens, and Stephen R Cole. Generalizing study results: a potential outcomes perspective. *Epidemiology (Cambridge, Mass.)*, 28(4):553, 2017.

Fan Li, Ashley L Buchanan, and Stephen R Cole. Generalizing trial evidence to target populations in non-nested designs: Applications to aids clinical trials. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(3):669–697, 2022.

Charles F Manski. Identification problems in the social sciences and everyday life. *Southern Economic Journal*, 70(1):11–21, 2003.

Charles F Manski. *Identification for prediction and decision*. Harvard University Press, 2007.

Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9 (4):403, 2004.

Allan H Murphy and Robert L Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 26(1):41–47, 1977.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI conference on artificial intelligence*, 29 (1), 2015.

Xinkun Nie, Guido Imbens, and Stefan Wager. Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. *arXiv preprint arXiv:2112.04723*, 2021.

Judea Pearl and Elias Bareinboim. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4):579 – 595, 2014. doi: 10.1214/14-STS486.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Sashank Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015.

Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

Claude M Setodji, Daniel F McCaffrey, Lane F Burgette, Daniel Almirall, and Beth Ann Griffin. The right tool for the job: Choosing between covariate balancing and generalized boosted model propensity scores. *Epidemiology (Cambridge, Mass.)*, 28(6):802, 2017.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Changyu Shen, Xiaochun Li, Lingling Li, and Martin C Were. Sensitivity analysis for causal inference using inverse probability weighting. *Biometrical journal*, 53(5):822–837, 2011.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Cedric M Smith. Origin and uses of primum non nocere—above all, do no harm! *The Journal of Clinical Pharmacology*, 45(4):371–377, 2005.

Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(2):369–386, 2011.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20, 2007.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Chunhao Tu. Comparison of various machine learning algorithms for estimating generalized propensity score. *Journal of Statistical Computation and Simulation*, 89(4):708–719, 2019.

Victor Veitch and Anisha Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Advances in neural information processing systems*, 33:10999–11009, 2020.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Lan Wang, Yu Zhou, Rui Song, and Ben Sherwood. Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254, 2018.

Daniel Westreich. *Epidemiology by Design: A Causal Approach to the Health Sciences*. Oxford University Press, Incorporated, 2019. ISBN 9780190665760.

Daniel Westreich, Jessie K Edwards, Catherine R Lesko, Elizabeth Stuart, and Stephen R Cole. Transportability of trial results using inverse odds of sampling weights. *American journal of epidemiology*, 186(8):1010–1014, 2017.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. *Advances in neural information processing systems*, 32, 2019.

Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):735–761, 2019.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

 https://doi.org/10.52202/079017-2230

Supplementary Material

In this supplementary material, we provide the proof outlined in Section A and additional details on the numerical experiments discussed in Section B.

A Proof

In this context, $\mathbb{P}$ represents the probability over samples drawn from both $p(X, U, A, L|S = 1)$ and $p_\pi(X, U, A, L|S = 0)$. The proof technique presented here builds upon several results established in Vovk et al. (2005); Tibshirani et al. (2019); Jin et al. (2023); Ek et al. (2023), and for completeness we present the proof in full.

Following Ek et al. (2023) we introduce β such that $0 < \beta < \alpha$. Use (12) to construct the limit

$$\ell_\alpha^\Gamma(\mathcal{D}) = \inf\left\{\ell : \widehat{F}(\ell; \mathcal{D}'', \overline{w}_\beta^\Gamma(\mathcal{D}')) \geq \frac{1-\alpha}{1-\beta}\right\}, \tag{16}$$

where $\overline{w}_\beta^\Gamma(\mathcal{D}')$ is defined in (10). We want to lower bound the probability of $L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D})$. Note that

$$\begin{aligned}\mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D})\} = \mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}) \mid \overline{W}_{n+1}^\Gamma \leq \overline{w}_\beta^\Gamma(\mathcal{D}')\}\, \mathbb{P}\{\overline{W}_{n+1}^\Gamma \leq \overline{w}_\beta^\Gamma(\mathcal{D}')\} \\ + \mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}) \mid \overline{W}_{n+1}^\Gamma > \overline{w}_\beta^\Gamma(\mathcal{D}')\}\, \mathbb{P}\{\, \overline{W}_{n+1}^\Gamma > \overline{w}_\beta^\Gamma(\mathcal{D}')\},\end{aligned}$$

from the law of total probability. The second term is a lower bounded by zero, and we have

$$\mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D})\} \geq \mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}) \mid \overline{W}_{n+1}^\Gamma \leq \overline{w}_\beta^\Gamma(\mathcal{D}')\}\, \mathbb{P}\{\overline{W}_{n+1}^\Gamma \leq \overline{w}_\beta^\Gamma(\mathcal{D}')\}. \tag{17}$$

Let us focus on the second factor in (17). From the construction in (10) the probability of $\overline{W}_{n+1}^\Gamma \leq \overline{w}_\beta^\Gamma(\mathcal{D}')$ is lower bounded by

$$\mathbb{P}\{\overline{W}_{n+1}^\Gamma \leq \overline{w}_\beta^\Gamma(\mathcal{D}')\} \geq 1 - \beta, \tag{18}$$

see Vovk et al. (2005); Lei et al. (2018, thm. 2.1).

We now proceed to bound the first factor in (17), i.e., $\mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}) \mid \overline{W}_{n+1}^\Gamma \leq \overline{w}_\beta^\Gamma(\mathcal{D}')\}$. Define the following limit

$$\ell_\alpha^\Gamma(\mathcal{D}'', \overline{W}_{n+1}^\Gamma) = \inf\left\{\ell : \widehat{F}(\ell; \mathcal{D}'', \overline{W}_{n+1}^\Gamma) \geq \frac{1-\alpha}{1-\beta}\right\}, \tag{19}$$

where $\overline{W}_{n+1}^\Gamma \geq W_{n+1}$ is given in (8). Comparing this limit with the one defined in (16), we see that

$$\begin{aligned}\mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}) \mid \overline{W}_{n+1}^\Gamma \leq \overline{w}_\beta^\Gamma(\mathcal{D}')\} &\geq \mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}'', \overline{W}_{n+1}^\Gamma) \mid \overline{W}_{n+1}^\Gamma \leq \overline{w}_\beta^\Gamma(\mathcal{D}')\} \\ &= \mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}'', \overline{W}_{n+1}^\Gamma)\}\,,\end{aligned}$$

whenever $\overline{W}_{n+1}^\Gamma \leq \overline{w}_\beta^\Gamma(\mathcal{D}')$. The second line results from applying sample splitting, which gurantess that $L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}'', \overline{W}_{n+1}^\Gamma)$ and $\overline{W}_{n+1}^\Gamma \leq \overline{w}_\beta^\Gamma(\mathcal{D}')$ are independent.

To lower bound $\mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}'', \overline{W}_{n+1}^\Gamma)\}$, we will make use of the following inequality,

$$\mathbb{E}\left[\widehat{F}(\ell_\alpha^\Gamma; \mathcal{D}'', \overline{W}_{n+1}^\Gamma)\right] = \mathbb{E}\left[\frac{\sum_{i\in\mathcal{D}''} \underline{W}_i^\Gamma \mathbb{1}(L_i \leq \ell_\alpha^\Gamma)}{\sum_{i\in\mathcal{D}''}\left[\underline{W}_i^\Gamma \mathbb{1}(L_i \leq \ell_\alpha^\Gamma) + \overline{W}_i^\Gamma \mathbb{1}(L_i > \ell_\alpha^\Gamma)\right] + \overline{W}_{n+1}^\Gamma}\right] \geq \frac{1-\alpha}{1-\beta}, \tag{20}$$

that holds by construction.

First, define $\mathcal{S}_+$ as an unordered set of the following elements

$$\big((X_{m'+1}, U_{m'+1}, A_{m'+1}, L_{m'+1}), \ldots, (X_m, U_m, A_m, L_m), (X_{n+1}, U_{n+1}, A_{n+1}, L_{n+1})\big).$$

From Tibshirani et al. (2019, thm. 2) we have that the out-of-sample loss L_{n+1} has the (conditional) cdf

$$\mathbb{P}\{L_{n+1} \leq \ell \mid \mathcal{S}_+\} = \sum_{i \in \mathcal{S}_+} p_i \mathbb{1}(\ell_i \leq \ell) = \frac{\sum_{i \in \mathcal{S}_+} w_i \mathbb{1}(L_i \leq \ell)}{\sum_{i \in \mathcal{S}_+} w_i}, \tag{21}$$

where w_i quantifies the distribution shift for sample i using the (unobservable) ratio (7). Next, we build on the proof method used in Jin et al. (2023, thm. 2.2). Use the limit $\ell_\alpha^\Gamma(\mathcal{D}', \overline{W}_{n+1}^\Gamma)$ from (19) in (21) and apply the law of total expectation to perform marginalization over $\mathcal{S}_+$

$$\begin{aligned}\mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}', \overline{W}_{n+1}^\Gamma)\} &= \mathbb{E}\big[\mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}', \overline{W}_{n+1}^\Gamma) \mid \mathcal{S}_+\}\big] \\ &= \mathbb{E}\left[\frac{\sum_{i \in \mathcal{S}_+} W_i \mathbb{1}(L_i \leq \ell_\alpha^\Gamma(\mathcal{D}', \overline{W}_{n+1}^\Gamma))}{\sum_{i \in \mathcal{S}_+} W_i}\right].\end{aligned} \tag{22}$$

We can now proceed to establish a lower bound for this probability. Combining (20) and (22), we have that

$$\begin{aligned}&\mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}', \overline{W}_{n+1}^\Gamma)\} - \frac{1-\alpha}{1-\beta} \\ &\geq \mathbb{E}\left[\frac{\sum_{i \in \mathcal{S}_+} W_i \mathbb{1}(L_i \leq \ell_\alpha^\Gamma)}{\sum_{i \in \mathcal{S}_+} W_i}\right] - \mathbb{E}\left[\frac{\sum_{i \in \mathcal{D}''} \underline{W}_i^\Gamma \mathbb{1}(L_i \leq \ell_\alpha^\Gamma)}{\sum_{i \in \mathcal{D}''} \left[\underline{W}_i^\Gamma \mathbb{1}(L_i \leq \ell_\alpha^\Gamma) + \overline{W}_i^\Gamma \mathbb{1}(L_i > \ell_\alpha^\Gamma)\right] + \overline{W}_{n+1}^\Gamma}\right] \\ &= \mathbb{E}\left[\frac{(*)}{\left[\sum_{i \in \mathcal{S}_+} W_i\right]\left[\sum_{i \in \mathcal{D}''} \left[\underline{W}_i^\Gamma \mathbb{1}(L_i \leq \ell_\alpha^\Gamma) + \overline{W}_i^\Gamma \mathbb{1}(L_i > \ell_\alpha^\Gamma)\right] + \overline{W}_{n+1}^\Gamma\right]}\right],\end{aligned}$$

where

$$\begin{aligned}(*) &= \left[\sum_{i \in \mathcal{S}_+} W_i \mathbb{1}(L_i \leq \ell_\alpha^\Gamma)\right]\left[\sum_{i \in \mathcal{D}''} \overline{W}_i^\Gamma \mathbb{1}(L_i > \ell_\alpha^\Gamma) + \overline{W}_{n+1}^\Gamma\right] \\ &\quad - \left[\sum_{i \in \mathcal{D}''} \underline{W}_i^\Gamma \mathbb{1}(L_i \leq \ell_\alpha^\Gamma)\right]\left[\sum_{i \in \mathcal{S}_+} W_i \mathbb{1}(L_i > \ell_\alpha^\Gamma)\right] \\ &\geq \left[\sum_{i \in \mathcal{D}''} W_i \mathbb{1}(L_i \leq \ell_\alpha^\Gamma)\right]\left[\sum_{i \in \mathcal{D}''} W_i \mathbb{1}(L_i > \ell_\alpha^\Gamma) + W_{n+1}\right] \\ &\quad - \left[\sum_{i \in \mathcal{D}''} W_i \mathbb{1}(L_i \leq \ell_\alpha^\Gamma)\right]\left[\sum_{i \in \mathcal{D}''} W_i \mathbb{1}(L_i > \ell_\alpha^\Gamma) + W_{n+1}\right] \\ &= 0.\end{aligned}$$

We use the bounds provided in (8) to derive the inequality. Hence, we arrive at a valid limit

$$\mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}', \overline{W}_{n+1}^\Gamma)\} \geq \frac{1-\alpha}{1-\beta}. \tag{23}$$

Finally combine (18) and (23) to get

$$\begin{aligned}\mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D})\} &\geq \mathbb{P}\{L_{n+1} \leq \ell_\alpha^\Gamma(\mathcal{D}', \overline{W}_{n+1}^\Gamma)\}\, \mathbb{P}\{\overline{W}_{n+1}^\Gamma \leq \overline{w}_\beta^\Gamma(\mathcal{D}')\} \\ &\geq \frac{1-\alpha}{1-\beta}(1-\beta) \\ &= 1-\alpha.\end{aligned} \tag{24}$$

We choose β as in (13) to get the tightest limit. As L_{n+1} is drawn from $p_\pi(X, U, A, L|S=0)$, we can express (24) as shown in (4) for the sake of notational clarity.

https://doi.org/10.52202/079017-2230

Table 2: Hyperparameters used for XGBoost in Section 5.1.

Parameter	Value
n_estimators	100
max_depth	2
learning_rate	0.05
objective	binary:logistic
min_child_weight	1
subsample	0.6
colsample_bytree	0.8
colsample_bylevel	0.4
scale_pos_weight	$n_{s=0}/n_{s=1}$

Table 3: Hyperparameters used for XGBoost in Section 5.2.

Parameter	Value
n_estimators	50
max_depth	1
learning_rate	0.2
objective	binary:logistic
min_child_weight	1
subsample	0.4
colsample_bytree	0.7
colsample_bylevel	0.8
scale_pos_weight	$n_{s=0}/n_{s=1}$

B Numerical Experiments

Additional information of the numerical experiments outlined in Section 5 can be found in this supplementary material and the code used for the experiments can be accessed here `https://github.com/sofiaek/policy-evaluation-rct`.

All experiments were conducted using Version 1.7 of the Python implementation of XGBoost (Apache-2.0 License). A comprehensive list of hyperparameters is available in Table 2 for the synthetic case and Table 3 for the NHANES case. The hyperparameters were selected through a random search involving 200 runs, employing 5-fold cross-validation with the F1 score as the optimization metric. All experiments were performed on a laptop with the following specifications: Intel Core i7-8650 CPU @ 1.9GHz, 16 GB DDR4 RAM, and Windows 10 Pro 64-bit operating system. The experiments utilized only the CPU. The total time required to run all the experiments was approximately half an hour.

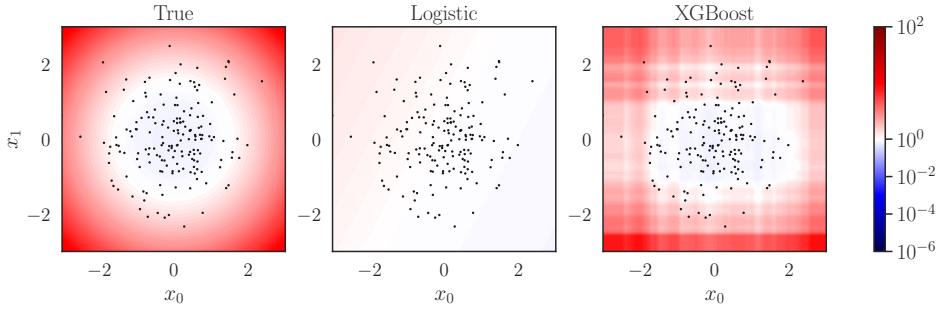
B.1 Synthetic data

We extend the experiments in Section 5.1 to include two extra target populations `C`, respective `D`. The full list of parameters used in (15) are given in Table 4.

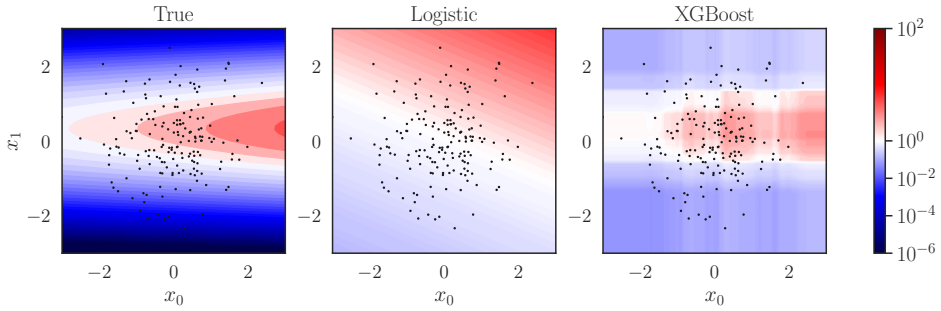
To illustrate the generality of the proposed methodology, we consider two different fitted models $\widehat{p}(S|X)$: a logistic model, which is conventionally used in the causal inference literature, and the more flexible tree-based ensemble model trained by XGBoost. Figure 7 compares nominal sampling odds obtained from the fitted models with the unknown odds, $p(S=0|X)/p(S=1|X)$, for the target populations `C`, respective `D`. This corresponds to a case without unmeasured selection factors U. In all these cases, the logistic model is misspecified and miscalibrated while the XGBoost model provides odds that resemble the true ones. Note that the proposed algorithm can handle any model if (3) is satisfied. However, a well-performing model is generally required to achieve a small Γ and for a meaningful benchmark of the upper value of Γ.

Table 4: Means and variances of covariate distribution $p(X, U|S)$ in (15).

Population	$\mu_{0,S}$	$\mu_{1,S}$	$\mu_{U,S}$	$\sigma^2_{0,S}$	$\sigma^2_{1,S}$	$\sigma^2_{U,S}$
`A` $(S=0)$	0.5	0.5	0.5	1.0	1.0	1.0
`B` $(S=0)$	0.0	0.5	0.0	1.25	1.5	1.25
`C` $(S=0)$	0.0	0.0	0.0	1.5	1.5	1.5
`D` $(S=0)$	0.25	0.25	0.25	1.0	0.25	0.5
`Trial` $(S=1)$	0.0	0.0	0.0	1.0	1.0	1.0



(a) Sampling odds for for target population `C`.



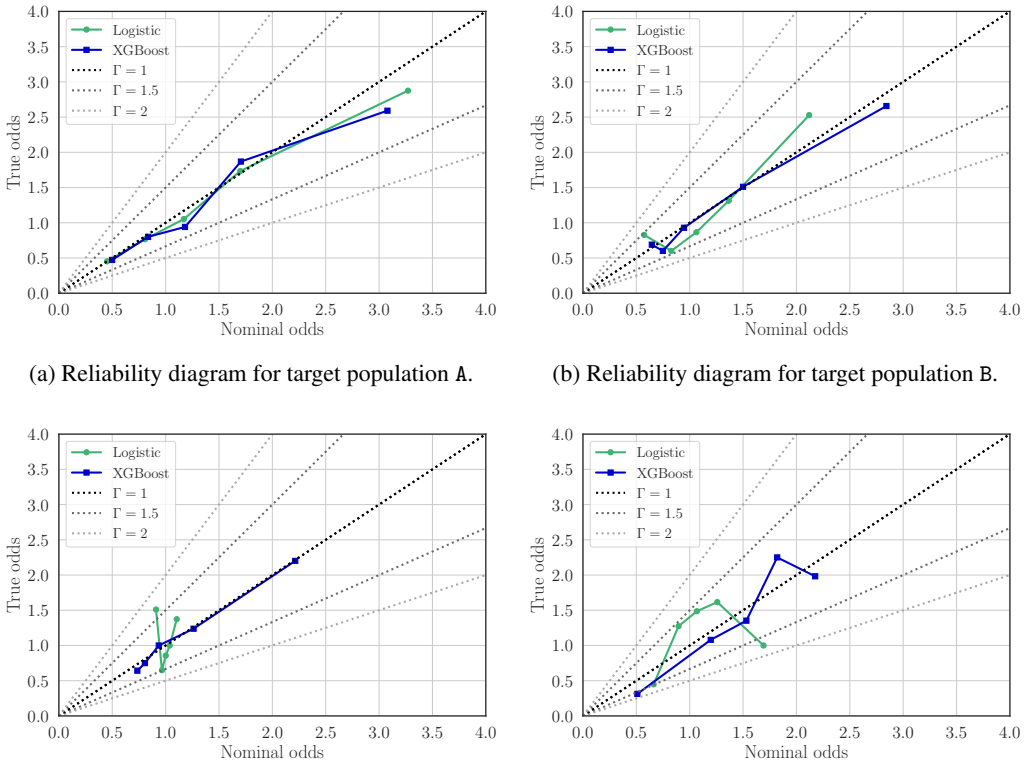
(b) Sampling odds for target population `D`.

Figure 7: Odds $p(S=0|X)/p(S=1|X)$ compared with nominal odds obtained from logistic and XGBoost models $\widehat{p}(S|X)$. The dots are a random subsample of the trial samples.

In Figure 8, we use the reliability diagram technique described in subsection 4.1 to assess and compare the performance of logistic and XGBoost models of the nominal odds, $\widehat{\text{odds}}(X)$, for the synthetic data. For all numerical experiments, we use 5 bins. For target population `A` in Figure 8a, both models are close to the diagonal and it is reasonable to believe that they are flexible enough to model the odds. For the other three populations, `B`, `C` and `D`, in Figure 8b, 8c respective 8d it is evident that the XGBoost model shows a closer alignment with the diagonal compared to the logistic model. This is in line with the results in Figure 5b and 7 where the XGBoost model is visually closer to the true model.

We now use the observed covariates to benchmark appropriate upper bounds for Γ for all target populations as described in subsection 4.1. Figure 9a shows the evaluation with respect to population `A`. Assuming that the unmeasured selection factors U are of similar strength as the weakest covariate, X_1, a Γ value of 2.5 could be a credible choice, as it covers 90% of the odds ratios. Figure 9b shows the same evaluation with respect to population `B`. A Γ value of 1.2 could be a credible choice for the logistic regression model and a Γ value of 2 could be appropriate for the XGBoost model. However, from Figure 5b we know that the logistic model is misspecified in this scenario. For population `C` in Figure 9c, 1 and 2.5 seem to be suitable Γ values for the logistic regression model respective the XGBoost model. Similarly, for population `D` in Figure 9d, 1.5 and 2.2 could be suitable values for the two models. The logistic model is again misspecified in populations `C` and `D`.

https://doi.org/10.52202/079017-2230



(a) Reliability diagram for target population A.

(b) Reliability diagram for target population B.

(c) Reliability diagram for target population C.

(d) Reliability diagram for target population D.

Figure 8: Reliability diagram of the observed odds against the average predicted nominal odds obtained from logistic and XGBoost-trained models $\widehat{p}(S|X)$.

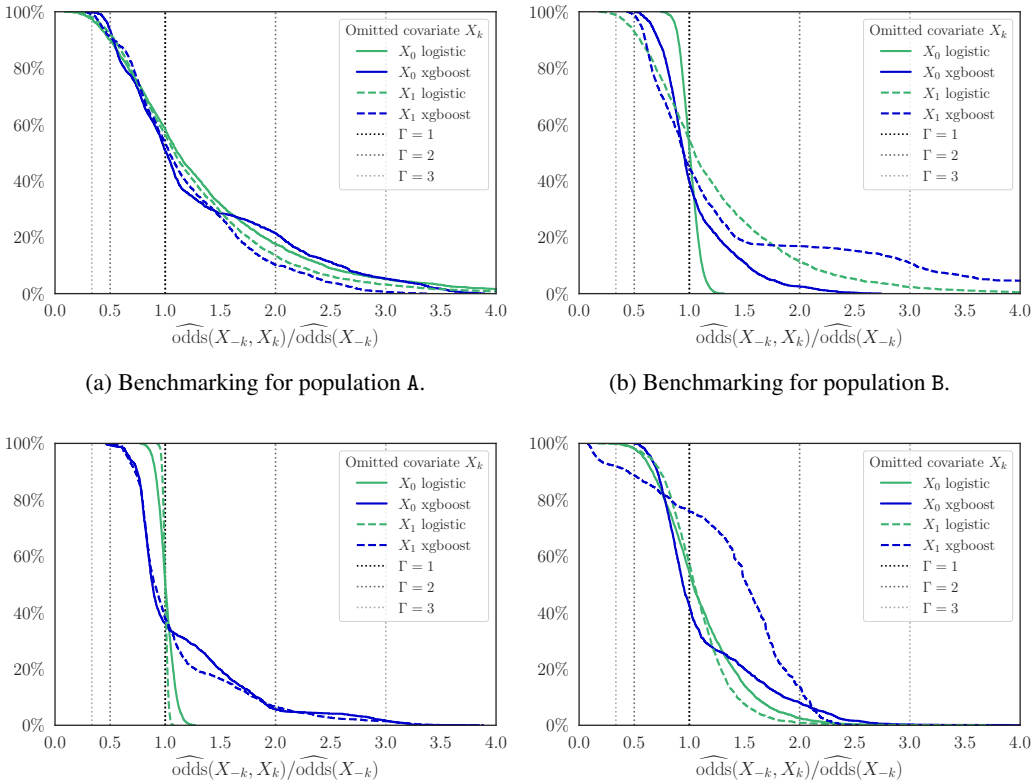
In Figure 10, we use the limit curves for the benchmark in (14) and the proposed method to evaluate the out-of-sample loss of the 'treat all' policy π_1, i.e.

$$p_{\pi_1}(A = 0 \mid X) = 1.$$

Figure 10a shows the evaluation with respect to population A. When $\Gamma = 1$ all the limit curves are similar. The curves for the proposed method also illustrate increasing the credibility of the models results in less informative inferences. However, their informativeness stays above 90% for odds miscalibration degrees $\Gamma \in [1, 2]$. We also evaluate the miscoverage gap of the curves and observe that the benchmark and the limit curves for $\Gamma = 1$ have a negative miscoverage gap. As the degree of miscalibration Γ increases to 2, the limit curves exhibit positive miscoverage gaps, where XGBoost results in slightly less conservative inferences than the logistic model does. We continue with population B and C in Figure 10b respective 10c. The limit curves derived for the baselines closely align with the curves modelled using logistic regression when $\Gamma = 1$, but is consistently lower than the curves modelled using XGBoost. For the certified curves the informativeness stays above 90% for odds miscalibration degrees $\Gamma \in [1, 2]$. For the miscoverage gap, the baselines and the limit curves for $\Gamma = 1$ are invalid. As the degree of miscalibration Γ increases to 2, all limit curves indicate positive miscoverage gaps. In Figure 10d, we evaluate the same for population D. The limit curves modelled using the baseline and logistic regression when $\Gamma = 1$ infer consistently higher losses than the curve modelled using XGBoost. For the curve using the logistic model the informativeness stays above 90% for odds miscalibration degrees $\Gamma \in [1, 2]$. The same figure for the XGBoost model is approximately 95%. For the miscoverage gap, the baseline and the limit curves for $\Gamma = 1$ are invalid. Again, as the degree of miscalibration Γ increases to 2, all limit curves indicate positive miscoverage gaps.

We now turn to comparing the 'treat all' policy with a 'treat none' policy, i.e.,

$$p_{\pi_0}(A = 0 \mid X) = 1,$$



(a) Benchmarking for population A.

(b) Benchmarking for population B.

(c) Benchmarking for population C.

(d) Benchmarking for population D.

Figure 9: Benchmarking the degree of miscalibration Γ using omitted covariates.

for population A. Their expected losses are estimated using (6) as $V^{\pi_0}_{\text{IPSW}} = 1.64$ and $V^{\pi_1}_{\text{IPSW}} = 1.54$, respectively. This evaluation suggests that π_1 is preferable to π_0. However, the limit curves presented in Figure 11a provide a more detailed picture in terms of out-of-sample losses: the tail losses certified for the 'treat none' policy π_0 are lower than those certified for π_1. This illustrates the cautionary principle built into the policy evaluations. Similar results are observed for target population D. While for target populations B and C, both $V^{\pi_0}_{\text{IPSW}}$ and $V^{\pi_1}_{\text{IPSW}}$ exhibit comparable sizes, the proposed limit curves still offer a more detailed understanding of out-of-sample losses.

For completeness, the actual degree of miscalibration for all target populations are visualized in Figure 12 for the case without unmeasured selection factors U.

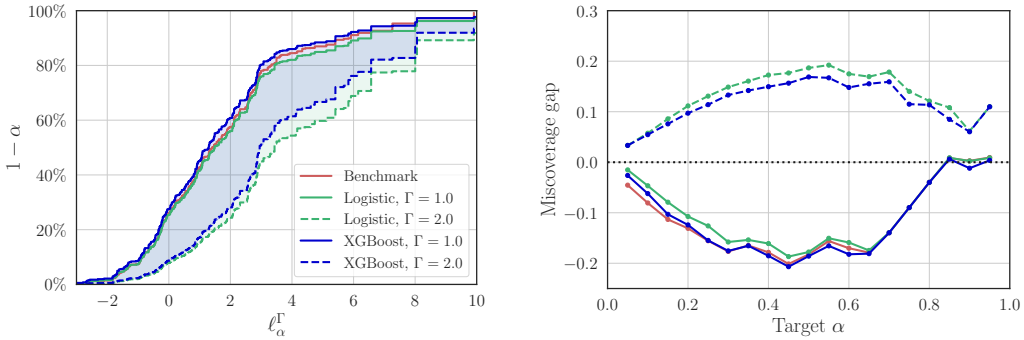
B.2 Seafood consumption policies

The National Health and Nutrition Examination Survey data (NHANES) is produced by federal agencies and is in the public domain, allowing it to be reproduced without permission. In our evaluation the data is split into observational data $\mathcal{D}_0$ and trial data $\mathcal{D}_1$ based on the covariates age, income, gender and smoking history (age and income are standardized), e.g.

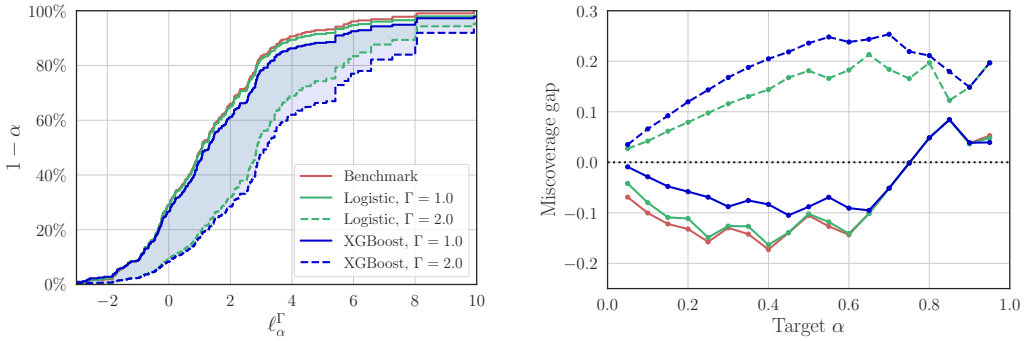
$$p(S = 1|X, U) = 0.25 \cdot [f(X_{\text{age}}) + f(X_{\text{income}})] + 0.05 \cdot [\mathbb{1}(X_{\text{male}} = 1) + \mathbb{1}(X_{\text{smoking ever}} = 1)] + 0.3,$$

where $f(\cdot)$ is the sigmoid function.

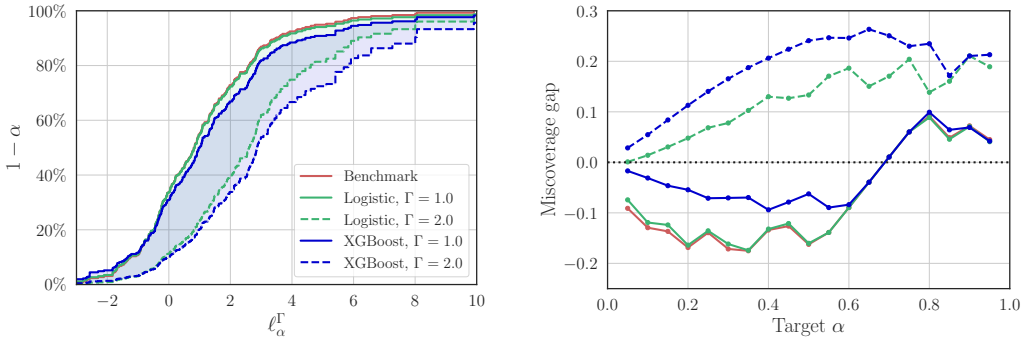
https://doi.org/10.52202/079017-2230



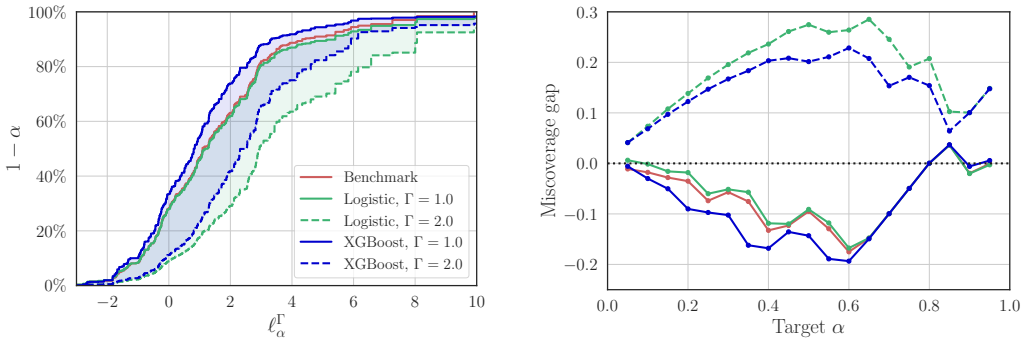
(a) Limit curve and miscoverage gap for target population A.



(b) Limit curve and miscoverage gap for target population B.

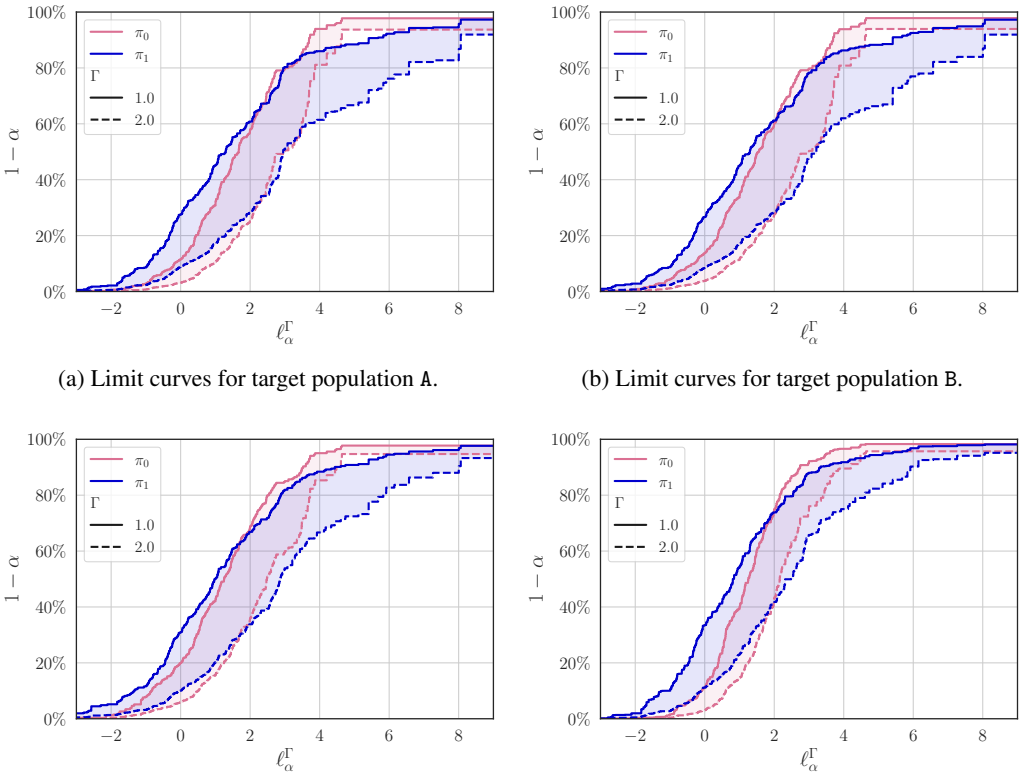


(c) Limit curve and miscoverage gap for target population C.



(d) Limit curve and miscoverage gap for target population D.

Figure 10: Evaluating 'treat all' policy π_1 for different target populations with degrees of miscalibration $\Gamma \in [1, 2]$.



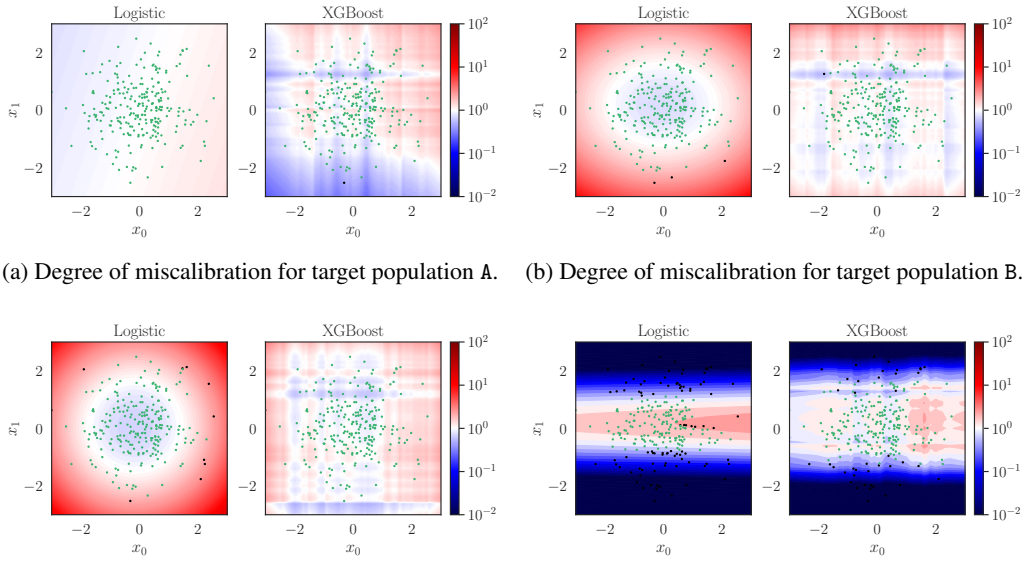
(a) Limit curves for target population A.

(b) Limit curves for target population B.

(c) Limit curves for target population C.

(d) Limit curves for target population D.

Figure 11: Limit curves for π_0 and π_1 for different target populations certified for $\Gamma \in [1, 2]$.



(a) Degree of miscalibration for target population A.

(b) Degree of miscalibration for target population B.

(c) Degree of miscalibration for target population C.

(d) Degree of miscalibration for target population D.

Figure 12: The actual degree of miscalibration for different target populations. The dots are a random subsample of the trial samples, where the green ones correspond to a degree of miscalibration within $\Gamma = 2$ and the black ones to a degree of miscalibration not bounded by $\Gamma = 2$.

https://doi.org/10.52202/079017-2230

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main assumptions are presented in Section 2, the method is detailed in Section 4, and the experiments are described in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The assumptions are included in Section 2 and some limitations are raised in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Main assumptions are available in Section 2, the method is described in Section 4 and the proof in the supplementary material A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experiments are described in Section 5, with more details in supplementary material B. The code is also provided in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-2230

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is available in the supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiments are described in Section 5, with more details in the supplementay material B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The miscoverage gap is reported, validating the statistical guarantee of the method.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Included in the supplementary material B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and this research conforms to its guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Included in Section 7.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/079017-2230

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Included in the supplementary material B.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

 https://doi.org/10.52202/079017-2230