Measuring Model Performance in the Presence of an Intervention

Winston Chen[1], **Michael W. Sjoding**[2], **Jenna Wiens**[1]

[1]Division of Computer Science and Engineering, University of Michigan, Ann Arbor, MI 48109, USA
[2]Division of Pulmonary and Critical Care Medicine, University of Michigan Medical School, Ann Arbor, MI 48109, USA
chenwt@umich.edu, msjoding@med.umich.edu, wiensj@umich.edu

Abstract

AI models are often evaluated based on their ability to predict the outcome of interest. However, in many AI for social impact applications, the presence of an intervention that affects the outcome can bias the evaluation. Randomized controlled trials (RCTs) randomly assign interventions, allowing data from the control group to be used for unbiased model evaluation. However, this approach is inefficient because it ignores data from the treatment group. Given the complexity and cost often associated with RCTs, making the most use of the data is essential. Thus, we investigate model evaluation strategies that leverage all data from an RCT. First, we theoretically quantify the estimation bias that arises from naïvely aggregating performance estimates from treatment and control groups and derive the condition under which this bias leads to incorrect model selection. Leveraging these theoretical insights, we propose nuisance parameter weighting (NPW), an unbiased model evaluation approach that reweights data from the treatment group to mimic the distribution of samples that would or would not experience the outcome under no intervention. Using synthetic and real-world datasets, we demonstrate that our proposed evaluation approach consistently yields better model selection than the standard approach, which ignores data from the treatment group, across various intervention effect and sample size settings. Our contribution represents a meaningful step towards more efficient model evaluation in real-world contexts.

Code — https://github.com/MLD3/NPW
Extended version — https://arxiv.org/abs/2511.05805

Introduction

Assessing a model's ability to predict the outcome of interest is critical before real-world deployment. However, in many applications studied by AI for social impact, the presence of an intervention that influences the outcome of interest can complicate model evaluation. For example, in healthcare, AI models have been developed to predict readmission (Yu et al. 2015; Huang et al. 2021). At the same time, interventions designed to prevent readmission (*e.g.,* post-discharge phone check-in) are used to reduce the overall readmission rate (Hansen et al. 2011). To allocate limited intervention resources, hospitals often use deterministic rules based on patients' risk index (e.g., the LACE index > 10) to decide who should receive the intervention (Teh and Janus 2018). Similar examples also exist in domains such as: (1) public health (Ayer et al. 2019; Deo et al. 2015; Amarasingham et al. 2013; Wang et al. 2023; Yang et al. 2020), (2) infrastructure maintenance (Gerum, Altay, and Baykal-Gürsoy 2019; Yeter, Garbatov, and Soares 2020), and (3) education support programs (Mac Iver et al. 2019; Adnan et al. 2021).

When measuring model performance in the presence of an intervention, using all data in the evaluation may introduce *outcome bias* (Pajouheshnia et al. 2017), because the intervention can alter the observed outcomes. Evaluating with un-intervened data (i.e., data in which the intervention was not applied) avoids outcome bias but may introduce *selection bias*, because interventions are often not assigned randomly. Inverse propensity weighting (IPW)-based methods that reweight data based on their likelihood of intervention can mitigate selection bias, but they require some randomness within the intervention assignment (e.g., based on probabilities) (Pajouheshnia et al. 2017; Coston et al. 2020; Boyer, Dahabreh, and Steingrimsson 2023; Keogh and Van Geloven 2024). In real-world applications, interventions are often assigned deterministically (e.g., based on thresholds), rendering IPW inapplicable.

Temporarily pausing the intervention for a period allows for unbiased model evaluation. Though conceptually simple, this approach can be challenging to operationalize in practice. Well-established interventions are often supported by a designated team. A temporary pause may require furloughing the team or appropriately redirecting their efforts. Instead, when feasible, a randomized controlled trial (RCT) could eliminate selection bias while maintaining the same level of intervention and, thus, the same personnel effort. Although RCTs are often conducted to estimate treatment effects, they also provide unbiased 'control' data critical in evaluating model performance in many applications where interventions are otherwise routinely delivered.

During an RCT, the study cohort is randomly assigned to the treatment or control group, and only those in the treatment group receive the intervention. As a result, both treatment and control groups are representative samples of the study cohort. Thus, evaluation using data from the control group yields an unbiased estimate of model performance.

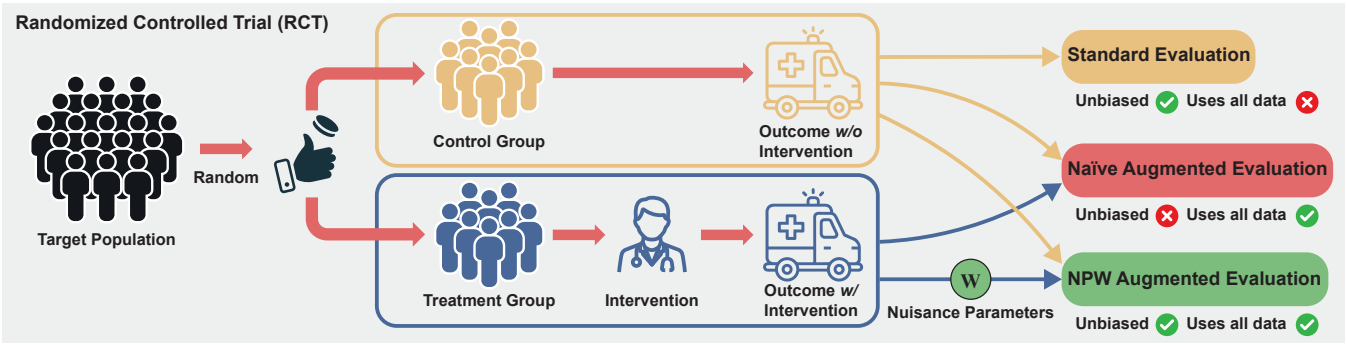

Figure 1: Overview of an RCT and different model evaluation approaches using RCT data. The standard evaluation is unbiased but only uses data from the control group. Naïve Augmented Evaluation uses data from both the control and treatment groups but is biased. Our proposed nuisance parameter weighting (NPW) augmented Evaluation is unbiased and uses all RCT data.

Conducting an RCT is often complex and expensive (Moore et al. 2020). Therefore, when such data are available, they should be fully leveraged. However, standard approaches to evaluating a predictive model with these data rely only on the control group, since the goal is to understand the predictive performance of a model in the absence of the intervention. This restriction reduces the effective sample size, increases variance in performance estimates, and potentially leads to inconclusive results.

In this work, we study how to augment standard evaluation with data from the treatment group in an RCT. We focus on the area under the receiver operating characteristic curve (AUROC), due to its wide adoption in AI model development and broad applicability to real-world applications (Ling, Huang, and Zhang 2003). However, our method can be generalized to any binary classification metric.

We start by considering naïve augmented AUROC, which aggregates AUROC estimates obtained separately from the treatment and control groups. We theoretically quantify its bias and derive an exact condition under which the bias leads to incorrect model selection. Building on our theoretical insights, we propose nuisance parameter weighting (NPW) augmented AUROC, an unbiased augmentation approach. NPW works by reweighting data from the treatment group to approximate the distribution of samples that would or would not experience the outcome under no intervention. We empirically show that NPW consistently leads to more accurate model selection across various intervention effect and sample size settings compared to the standard approach.

In summary, our contributions are:

- To our knowledge, we are the first to study measuring AUROC using data in which an intervention is applied,
- we derive the bias of naïve augmented AUROC and characterize conditions in which it selects incorrect models,
- we propose NPW, an unbiased approach to evaluate models using data from both control and treatment groups.
- we empirically demonstrate the advantages of our proposed approach in improving model selection.

Background

Problem Statement

Consider a target cohort characterized by $X \in \mathbb{R}^p$. Each individual is randomly assigned to the treatment ($T = 1$) or control ($T = 0$) group with probability π: $T \sim \text{Bern}(\pi)$. Each individual may experience a binary outcome, Y, sampled from a distribution conditioned on their T and X:

$$Y \mid X, T \sim \text{Bern}(\omega(X) + T\tau(X)) \tag{1}$$

where $\text{Bern}(\cdot)$ is the Bernoulli distribution, $\omega(\cdot)$ is the baseline outcome probability, and $\tau : X \to (-\omega(X), 1-\omega(X))$ represents the treatment effect conditional on X. The range of τ ensures that $\mathbf{P}(Y \mid X, T) \in (0, 1)$.

Given a dataset $\mathbb{D} = \{(x_i, y_i, t_i)\}_{1 \le i \le n}$ sampled from the above data generation process (DGP), we aim to evaluate a prediction model, $f : X \to \mathbb{R}$, in its AUROC for predicting the outcome under no-intervention, $\mathbb{E}\left[\text{AUC}(\mathbb{X}_0^+, \mathbb{X}_0^-, f)\right]$. The expectation is over $\mathbb{X}_0^+$ and $\mathbb{X}_0^-$, control samples that did or did not experience the outcome. They are defined as follows:

$$\mathbb{X}_0^+ = \{x_i | t_i = 0, y_i = 1\}, \quad \mathbb{X}_0^- = \{x_i | t_i = 0, y_i = 0\}.$$

The $\text{AUC}(\cdot, \cdot, \cdot)$ notation is defined as:

$$\text{AUC}(\mathbb{X}^+, \mathbb{X}^-, f) = \frac{1}{|\mathbb{X}^+||\mathbb{X}^-|} \sum_{\substack{x^{(i)} \in \mathbb{X}^+ \\ x^{(j)} \in \mathbb{X}^-}} \mathbb{1}_{\{f(x^{(i)}) > f(x^{(j)})\}},$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function, $\mathbb{X}^+$ and $\mathbb{X}^-$ both denote samples used to compute AUROC, but samples in $\mathbb{X}^+$ have experienced the outcome while samples in $\mathbb{X}^-$ have not.

Preliminaries

To estimate model f's AUROC under no intervention, simply using all data in $\mathbb{D}$ leads to biased estimates, because it produces a weighted average between the target AUROC ($\mathbb{E}\left[\text{AUC}(\mathbb{X}_0^+, \mathbb{X}_0^-, f)\right]$) and three other variant AUROCs.

To see this, we first define the following sets of data based on intervention and outcome variables:

$$\mathbb{D}_0 = \{(x_i, y_i) | t_i = 0\} \qquad \mathbb{D}_1 = \{(x_i, y_i) | t_i = 1\}$$
$$\mathbb{X}_0^- = \{x_i | t_i = 0, y_i = 0\} \quad \mathbb{X}_1^- = \{x_i | t_i = 1, y_i = 0\}$$
$$\mathbb{X}_0^+ = \{x_i | t_i = 0, y_i = 1\} \quad \mathbb{X}_1^+ = \{x_i | t_i = 1, y_i = 1\}$$

Then, we define short-hand notations for AUROCs estimated with only $\mathbb{D}_0$ and $\mathbb{D}_1$, respectively:

$$\text{AUC}_{\mathbb{D}_0}(f) = \text{AUC}(\mathbb{X}_0^+, \mathbb{X}_0^-, f)$$
$$\text{AUC}_{\mathbb{D}_1}(f) = \text{AUC}(\mathbb{X}_1^+, \mathbb{X}_1^-, f)$$

Leveraging this notation, the expected AUROC using all data can be expressed as:

$$\mathbb{E}\left[\text{AUC}_{\text{all}}(f)\right] = \mathbb{E}\Bigg[(1-\pi)^2 \underbrace{\text{AUC}_{\mathbb{D}_0}(f)}_{\text{unbiased term}} + \pi^2 \underbrace{\text{AUC}_{\mathbb{D}_1}(f)}_{\text{biased term 1}}$$
$$+(\pi - \pi^2) \underbrace{\left(\text{AUC}(\mathbb{X}_0^+, \mathbb{X}_1^-, f) + \text{AUC}(\mathbb{X}_1^+, \mathbb{X}_0^-, f)\right)}_{\text{biased term 2}}\Bigg]$$

Intuitively, the first term in the expression is unbiased, while the other terms introduce bias. We refer readers to Appendix Section 2 in the extended version of this paper for a detailed derivation of the above expression.

To obtain unbiased estimates of model f's AUROC, the standard approach only uses $\mathbb{D}_0$: $\text{AUC}_{\text{std}}(f) = \text{AUC}_{\mathbb{D}_0}(f)$. However, it ignores $\mathbb{D}_1$ entirely, and as a result, the estimate may be associated with high variance when the sample size is small. To augment standard AUROC estimates with $\mathbb{D}_1$, a naïve approach averages $\text{AUC}_{\mathbb{D}_0}$ and $\text{AUC}_{\mathbb{D}_1}$ with the randomization probability, π:

$$\text{AUC}_{\text{naïve}}(f) = (1-\pi)\text{AUC}_{\mathbb{D}_0}(f) + \pi\text{AUC}_{\mathbb{D}_1}(f) \quad (2)$$

The naïve augmented AUROC is biased due to the presence of $\text{AUC}_{\mathbb{D}_1}(f)$. However, compared to $\text{AUC}_{\text{all}}(f)$, it has only one bias term, allowing our theoretical analysis to succinctly quantify its bias and characterize scenarios where it incorrectly selects the model as the ground truth AUROC.

Theoretical Analysis of Naïve Augmentation

Naïvely augmenting the standard AUROC with data from the treatment group leads to biased estimates because the presence of intervention alters the outcomes. In this section, we formally derive its theoretical bias.

Bias of Naïve Augmented AUROC

Theorem 1 (Bias of Naïve Augmented AUROC (Eq. 2))**.** *Let μ_0 and μ_1 be the expected outcome for the control and treatment group, and τ be the average treatment effect (ATE):*

$$\mu_0 = \mathbb{E}_X[Y \mid X, T=0] = \mathbb{E}_X[\omega(X)]$$
$$\mu_1 = \mathbb{E}_X[Y \mid X, T=1] = \mathbb{E}_X[\omega(X) + \tau(X)]$$
$$\tau = \mathbb{E}_X[\tau(X)] = \mu_1 - \mu_0$$

Under our assumed DGP, the bias of $AUC_{naïve}(f)$ is:

$$\textit{Bias}\left(\textit{AUC}_{\textit{naïve}}(f)\right) = \alpha\delta(f) - \beta\sigma(f) \quad (3)$$

where α and β are bounded problem-specific real numbers defined as follows:

$$\alpha = \frac{\pi\tau(1-\mu_0-\mu_1)}{\mu_1(1-\mu_1)} < 1 \quad \beta = \frac{\pi}{\mu_1(1-\mu_1)} > 0$$

$\delta(f)$ represents model f's true AUROC improvement over a random prediction with AUROC being 0.5. $\sigma(f)$ denotes the covariance between the cumulative distribution function (CDF) of model f's prediction and the true individual-level effect, $\tau(X)$, also known as the conditional average treatment effect (CATE), of the intervention.

At a high level, our proof rewrites the bias definition of naïve augmented AUROC using our assumed DGP and Bayes' rule to obtain the simplified bias term in Eq. 3. Because it does not leverage AUROC-specific techniques, it is generalizable to any binary classification metric. Detailed derivation of Theorem 1 is provided in Appendix Section 1.

Theorem 1 shows that the bias of naïve augmented AUROC depends on a linear combination of: (1) the model's true AUROC: $\delta(f)$ and (2) the correlation between model and the CATE function: $\sigma(f)$. The exact contribution of each factor depends on problem-specific parameters: α and β.

Model Selection with Naïve Augmented AUROC

To contextualize the potential harm of this bias and better understand how bias in model evaluation could affect downstream decisions in real-world applications, we consider model selection using the estimated AUROC. We derive an exact condition under which model selection using naïve augmented AUROC leads to an incorrect model.

Theorem 2 (Condition for Incorrect Model Selection with Naïve Augmented AUROC)**.** *Given models f_1 and f_2, let $\theta(f)$ and $\hat{\theta}(\cdot)$ be models' true AUROC and expected AUROC estimate using naïve augmented evaluation, respectively. Without loss of generality, we consider $\hat{\theta}(f_1) > \hat{\theta}(f_2)$. When $\hat{\theta}(f_1) - \hat{\theta}(f_2) > \beta(\delta(f_1) - \delta(f_2))$, we will always have $\theta(f_1) < \theta(f_2)$, therefore $\hat{\theta}(\cdot)$ selects the incorrect model.*

Proof. Our proof starts by expressing the expected AUROC estimate from naïve augmentation in terms of its bias:

$$\hat{\theta}(f) = \theta(f) - \text{Bias}\left(\hat{\theta}(f)\right)$$

Applying Theorem 1 and the definition of $\delta(f)$, we get:

$$\hat{\theta}(f) = (1-\alpha)\,\theta(f) + \alpha 0.5 + \beta\sigma(f)$$

Given the above rewrite of $\hat{\theta}(f_1)$, we express the difference between $\theta(f_1)$ and $\theta(f_2)$ as:

$$\theta(f_1) - \theta(f_2) = \frac{(\hat{\theta}(f_1) - \hat{\theta}(f_2)) - \beta(\sigma(f_1) - \sigma(f_2))}{1-\alpha}$$

Because by definition $\alpha < 1$, the denominator of the right-hand side is positive. $\theta(f_1) < \theta(f_2)$ holds when

$$\hat{\theta}(f_1) - \hat{\theta}(f_2) < \beta\left(\sigma(f_1) - \sigma(f_2)\right)$$

Intuitively, we can think of this condition as two scenarios. In the first scenario, if the model selected by the naïve augmented AUROC has a lower or equal correlation with CATE than the model not selected, then naïve augmented AUROC will *never* choose the wrong model. This is because the right-hand side of the condition remains non-positive, while the left-hand side is always positive. In the second

scenario, the model selected by the naïve augmented AUROC correlates more with CATE than the model not selected. The naïve augmented AUROC selects the incorrect model if the estimated AUROC difference between the two models is smaller than a scaled version of their difference in CATE correlation. The scaling factor, β, is problem-specific but is positive and unbounded. Therefore, a large β in some scenarios can make the incorrect selection unavoidable as the estimated AUROC difference is upper bounded by 1.

Unbiased Augmentation with NPW

Here, we propose **N**uisance **P**arameter **W**eighting (NPW) augmented AUROC that leads to unbiased estimates while leveraging data from both control and treatment groups.

Similar to naïve augmentation, NPW estimates AUROC by averaging the standard estimate with an alternative estimate produced based on data from the treatment group, $\mathbb{D}_1$:

$$\text{AUC}_{\text{NPW}}(f) = (1-\pi)\text{AUC}_{\mathbb{D}_0}(f) + \pi\text{AUC}_{\text{alt}}(f)$$

The key to produce an unbiased $\text{AUC}_{\text{alt}}(f)$ lies in recovering distributions of samples that would or would not experience the outcome *without* intervention, denoted by $\mathbf{P}(X_0^+)$ and $\mathbf{P}(X_0^-)$. Because $\mathbb{D}_1$ is sampled from distributions of data that would or would not experience the outcome *with* intervention, denoted by $\mathbf{P}(X_1^+)$ and $\mathbf{P}(X_1^-)$, naïvely computing $\text{AUC}_{\text{alt}}(f)$ with $\mathbb{D}_1$ leads to biased estimates.

Our proposed NPW uses two weighting approaches to recover $\mathbf{P}(X_0^+)$ and $\mathbf{P}(X_0^-)$. The first approach uses data from the control group, $\mathbb{D}_0$, to learn a model that estimates the probability of a sample experiencing the outcome without intervention. Then, it re-weights $\mathbb{D}_1$ based on this probability to recover $\mathbf{P}(X_0^+)$ and $\mathbf{P}(X_0^-)$. However, this approach ignores the observed outcomes in $\mathbb{D}_1$, which provides information about $\mathbf{P}(X_1^+)$ and $\mathbf{P}(X_1^-)$. To leverage this information, we design a second approach to recover $\mathbf{P}(X_0^-)$ and $\mathbf{P}(X_0^+)$ by correcting for the effect or intervention in $\mathbf{P}(X_1^-)$ and $\mathbf{P}(X_1^+)$ using estimates of CATE.

The following section provides high-level derivations of the two weighting approaches. Additional details are available in Appendix Section 3.

Derivations of NPW

To derive the **first approach**, we write $\mathbf{P}(X_0^-)$ and $\mathbf{P}(X_0^+)$ using the Bayes' rule and the assumed DGP, leading to the following formulations in terms of weighted $\mathbf{P}(X)$, the natural distribution of samples in the target cohort:

$$\mathbf{P}(X_0^-) = \frac{1-\omega(X)}{1-\mu_0}\mathbf{P}(X),\ \mathbf{P}(X_0^+) = \frac{\omega(X)}{\mu_0}\mathbf{P}(X) \quad (4)$$

where the weights are based on $\omega(X)$ and μ_0, the individual and average probability of experiencing the outcome without intervention. Equation 4 shows that we can recover $\mathbf{P}(X_0^+)$ or $\mathbf{P}(X_0^-)$ by up-weighting samples with higher or lower probability of experiencing the outcome, receptively.

Because the intervention is random in RCT, $\mathbb{D}_1$ contains random samples from the target cohort. Therefore, we can approximate $\mathbf{P}(X_0^-)$ and $\mathbf{P}(X_0^+)$ by re-weighting $\mathbb{D}_1$ with $\hat{\omega}(X)$, an unbiased estimate of $\omega(X)$, following Equation 4. Formally we denote the AUROC estimated with this approach as: $\text{AUC}_{\hat{\omega}}(f) = \text{AUC}(\mathbb{X}_{\hat{\omega}}^+, \mathbb{X}_{\hat{\omega}}^-, f)$, where $\mathbb{X}_{\hat{\omega}}^+$ and $\mathbb{X}_{\hat{\omega}}^-$ are $\mathbb{D}_1$ reweighted to approximate $\mathbf{P}(X_0^+)$ and $\mathbf{P}(X_0^-)$.

To derive the **second approach**, we similarly apply the Bayes' rule and the DGP to rewrite $\mathbf{P}(X_0^-)$ and $\mathbf{P}(X_0^+)$ in terms of $\mathbf{P}(X_1^-)$ and $\mathbf{P}(X_1^+)$ as follows:

$$\begin{aligned}\mathbf{P}(X_0^-) &= \frac{1-\mu_1}{1-\mu_0}\mathbf{P}(X_1^-) + \frac{\tau(X)}{1-\mu_0}\mathbf{P}(X)\\ \mathbf{P}(X_0^+) &= \frac{\mu_1}{\mu_0}\mathbf{P}(X_1^+) - \frac{\tau(X)}{\mu_0}\mathbf{P}(X)\end{aligned} \quad (5)$$

This formulation leads to AUROC estimates with $\hat{\tau}(X)$, an unbiased estimate of $\tau(X)$: $\text{AUC}_{\hat{\tau}}(f) = \text{AUC}(\mathbb{X}_{\hat{\tau}}^+, \mathbb{X}_{\hat{\tau}}^-, f)$, where $\mathbb{X}_{\hat{\tau}}^+$ and $\mathbb{X}_{\hat{\tau}}^-$ are $\mathbb{D}_1$ reweighted following Equation 5.

The second approach recovers $\mathbf{P}(X_0^-)$ by up-weighting samples in $\mathbb{D}_1$ with higher CATE, because they are more likely not to experience the outcome without intervention. Conversely, to recover $\mathbf{P}(X_0^+)$, it down-weights higher CATE samples in $\mathbb{D}_1$ to account for the possibility that they experienced the outcome primarily due to the intervention.

Since $\text{AUC}_{\hat{\omega}}(f)$ and $\text{AUC}_{\hat{\tau}}(f)$ each depend on a nuisance parameter estimate that could have high variance in practice, we average them to further reduce the estimation variance. The **final form** of the alternative estimate in NPW is:

$$\text{AUC}_{\text{alt}}(f) = \frac{\text{AUC}_{\hat{\omega}}(f) + \text{AUC}_{\hat{\tau}}(f)}{2}$$

Practical Considerations

Calculating $\text{AUC}_{\hat{\omega}}(f)$ and $\text{AUC}_{\hat{\tau}}(f)$ with weighted sampling can be inefficient, as $\text{AUC}_{\hat{\tau}}(f)$ involves aggregating two distributions through their samples. Therefore, we implement them with weighted AUROC, which is equivalent to weighted sampling. See Appendix Section 3 for the exact formulation of weighted AUROC.

Experimental Setup

Does NPW empirically improve model evaluation and, in turn, model selection? In this section, we describe our experimental setup for investigating this question.

Datasets

We design a synthetic dataset to validate our theoretical results and evaluate NPW under controlled settings. Leveraging two real-world datasets, we evaluate NPW in realistic settings, where the nuisance parameters must be estimated.

The **synthetic** dataset is generated as follows:

$$\begin{aligned}&x_i \sim \mathcal{N}(0, \mathbf{I}_{20}), \quad t_i \sim \text{Bern}(0.5),\\ &\tau_i \mid x_i = \frac{\sigma(w_\tau x_i)\big(1 - w_y x_i\big)\Delta}{\frac{1}{n}\sum_{j=1}^{n}\sigma(w_\tau x_j)\big(1 - w_y x_j\big)},\\ &y_i \mid t_i, x_i \sim \text{Bern}(\sigma(w_y x_i) + t_i\tau_i)\,,\end{aligned}$$

where $w_y \in \mathbb{R}^{20}$ has 40% of entries drawn from $\mathcal{N}(0,1)$ and the rest set to 0, and $w_\tau \in \mathbb{R}^{20}$ is sampled from $\{0, 0.1, 0.2, 0.3, 0.4\}$ with probabilities

$\{0.8, 0.05, 0.05, 0.05, 0.05\}$. The scalar $\Delta \in [-1, 1]$ controls the average treatment effect (ATE). We generate 100,000 samples in total, using the AUROC computed on all control samples as the ground truth, and subsample $n = 200$ points to mimic the limited size of an RCT.

We use **AMR-UTI** dataset ($n = 15,806$) to simulate an RCT with real-world data. It is a dataset containing antimicrobial resistance results for patients with urinary tract infections (UTIs) (Oberst et al. 2020). AMR-UTI covers four types of antimicrobial treatments and uniquely provides ground-truth antimicrobial resistance labels for each treatment obtained via microbiology testing. We transform it into an RCT dataset by considering the two most common treatments: nitrofurantoin (NIT) and trimethoprim-sulfamethoxazole (SXT) as the treatment and control, respectively. Patients are randomly assigned to the treatment or control group with a probability of 0.5. The antimicrobial resistance labels are considered the outcome of interest. Since each pathogen associated with patient infection is tested for resistance to each antimicrobial, we have ground truth outcomes under both treatment and control settings. In our resampled RCT dataset, we observe the NIT resistance label for patients in the treatment group and the SXT resistance label for patients in the control group.

Lastly, we leverage the **Readmission** dataset, a real-world RCT dataset ($n = 1,518$), collected from Michigan Medicine (MM), a large academic medical center associated with the University of Michigan, to evaluate our approach with real RCT data. In this dataset, discharged patients are randomly assigned to the treatment or control group. Patients in the treatment group receive post-discharge phone check-in calls as an intervention. The outcome of interest is whether a patient experiences an unplanned readmission to the MM within 30 calendar days of discharge. Given that the study randomizes an existing intervention already in use at MM, the University of Michigan institutional review board (IRB) determined this RCT was a quality improvement (QI) initiative exempt from human subjects research, waiving the need for patient consent. One of the goals of this RCT is to determine whether the hospital should adopt a newly developed Epic readmission risk prediction model (Hwang et al. 2021) in its risk stratification workflow. The decision primarily depends on whether the Epic model significantly outperforms the LACE index (Van Walraven et al. 2010), a model currently used by the hospital to estimate readmission risk. Given the difficulty of predicting readmission, we expect only modest performance gains from the Epic model. Thus, standard evaluation could require a large sample size for a significant comparison. Our experiments evaluate different AUROC estimates using a statistical hypothesis testing framework, measuring whether they can reject the null hypothesis of Epic *does not* improve over LACE with limited samples. We provide additional details about this dataset in Appendix Section 5.

Models

To mimic a model selection scenario, we needed to generate several models to select based on their AUROC estimates. For the Synthetic and AMR-UTI datasets, we use gradient-boosted decision trees (Friedman 2001) to create models of varying performance. We trained each model with a varying number of samples, ranging from 100 to 1500 (for synthetic), and from 10 to $5,806$ (for AMR-UTI). To ensure a uniform distribution of models' true AUROC, we randomly sample one model from each 0.005 increment in true AUROC. This sampling procedure results in 65 models in the synthetic dataset and 31 in the AMR-UTI dataset.

For the Readmission dataset, we use the prospective predictions from the Epic model and LACE index during the RCT for evaluation; therefore, only existing models are applied, and no model training is necessary.

Nuisance Parameter Estimation

Our proposed NPW augmented AUROC requires estimating nuisance parameters, $\omega(X)$ and $\tau(X)$, to produce AUROC estimates. The nuisance parameters are directly available in the synthetic dataset due to the known DGP. Therefore, we simulate unbiased nuisance parameter estimates by adding Gaussian noise with variance v to the true nuisance parameters. We estimate nuisance parameters with the cross-fitting approach for the AMR-UTI and Readmission datasets. Specifically, when evaluating models with n data points, the nuisance parameters are also estimated with the same n data points. Data are split into k folds. For each fold, we use the remaining $k - 1$ folds to fit a nuisance parameter model with gradient-boosted decision trees, which then predicts nuisance parameters for the current fold.

Evaluation Metrics

In the synthetic dataset, we evaluate the quality of the AUROC estimate using the mean absolute error (MAE). For the AMR-UTI dataset, we evaluate whether AUROC-based model rankings align with the true rankings using the concordance index (C-index), which quantifies their agreement.

In the readmission dataset, we assess AUROC estimates by their statistical power to detect a true performance improvement between Epic and LACE models. Power is defined as the proportion of bootstrap-derived P-values below the significance level $\alpha = 0.05$. Each P-value is estimated from 1,000 bootstrap samples by comparing the AUROCs of LACE and Epic and calculating the proportion of samples where LACE outperforms Epic. We report mean values with 95% confidence intervals to summarize each metric.

Baselines

We evaluate two baselines to contextualize the performance of our proposed approach. (1) **Standard approach**: estimating AUROC with RCT data from the control group only. (2) **Naïve approach**: naïvely averaging AUROC estimated with RCT data from both the control and treatment groups.

Empirical Results

We evaluate NPW augmented AUROC in terms of three aspects compared to the baseline approaches: (1) Does NPW reduce estimation error? (2) Does NPW improve model selection performance? (3) Does NPW boost the statistical power of hypothesis testing in a real-world RCT?

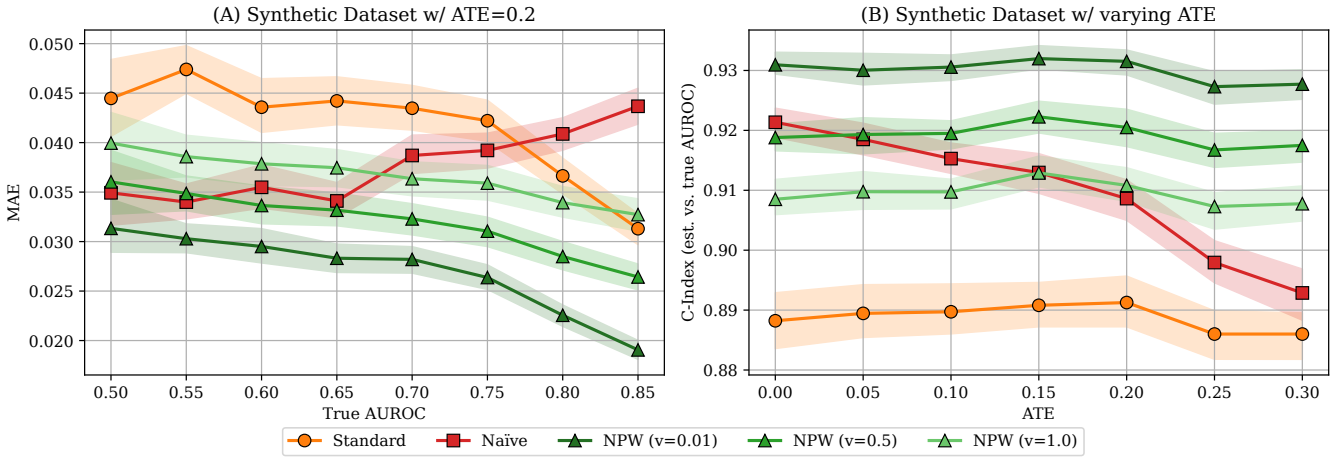

Figure 2: Empirical results with the synthetic dataset. (A) MAEs of different AUROC estimates for models with various ground true AUROCs. (B) C-index of model rankings induced by different AUROC estimates evaluated under interventions of various ATEs. In both figures, NPW consistently outperforms the standard approach, and its advantage over the standard approach increases as the variance (v) of nuisance parameter estimates decreases. Error bars are bootstrapped 95% confidence intervals.

NPW Augmentation Reduces Estimation Error

We start by evaluating the NPW and baselines in terms of their MAE. We use a synthetic dataset with ATE = 0.2 to compare all approaches in evaluating models with various true AUROCs. Because NPW's performance depends on the nuisance parameter estimates, we also simulate estimates with varying qualities by controlling their variances (v), with lower v indicating higher-quality estimates.

As shown in Figure 2(A), when the quality of nuisance parameter estimates is high ($v = 0.01$), NPW has lower average MAE than both the standard and naïve approaches across all evaluated models. As v increases, NPW's MAE worsens due to increased variance in the estimation. However, when the quality of nuisance parameter estimates is poor ($s = 1.0$), NPW still outperforms the standard approach across most models and beats the naïve approach on models with higher true AUROC (i.e., true AUROC ≥ 0.7).

Because NPW augmented AUROC removes the intervention effect on the treated data and is unbiased, we expect it to be consistent with different intervention ATEs. Additional results with other ATEs can be found in the Appendix Section 6.1. The results show similar trends to Figure 2(A).

Our results also show that naïve approach only reduces MAE compared to the standard approach when the intervention's ATE or model's true AUROC is low, and hurts the performance when either quantity is high, confirming our theory and highlighting the importance of unbiased augmentation with NPW. See Appendix Section 6.1 for details.

NPW Augmentation Improves Model Ranking

AUROC estimates are commonly used to rank models and select the best-performing one. Do AUROC estimates with lower MAE lead to improved model ranking performance? Here, we evaluate different AUROC estimation approaches in terms of the C-index of their induced model rankings against those determined by the models' true AUROCs. A higher C-index indicates better model ranking performance.

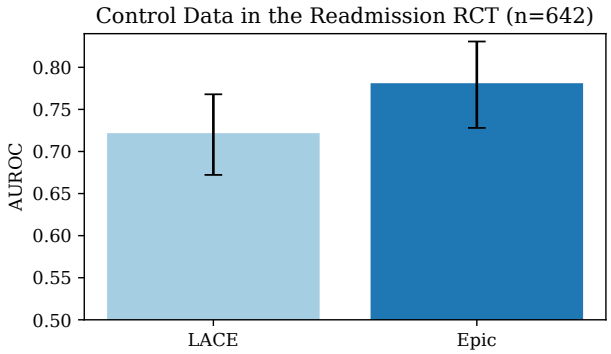

Figure 3: LACE and Epic's readmission prediction performance in terms of the AUROC, estimated using all control data. Error bars are bootstrapped 95% confidence intervals.

As shown in Figure 2(B), NPW augmented AUROCs consistently improve the C-index compared to the standard AUROC across a wide range of intervention ATEs. The improvement is robust against degradations in the quality of nuisance parameter estimates. Similar to Figure 2(A), NPW's C-index improvement over the standard AUROC diminishes as v increases. With high-quality nuisance parameter estimates ($v = 0.01$), NPW outperforms the naïve approach across all ATEs. Even with poor nuisance parameter estimates ($v = 1.0$), NPW remains advantageous over the naïve approach in settings with higher ATEs (ATE> 0.15). This is due to naïve approach's increasing estimation bias as the ATE increases, confirming our theoretical results.

In the AMR-UTI dataset, the nuisance parameters are estimated using the same samples for estimating AUROC. As

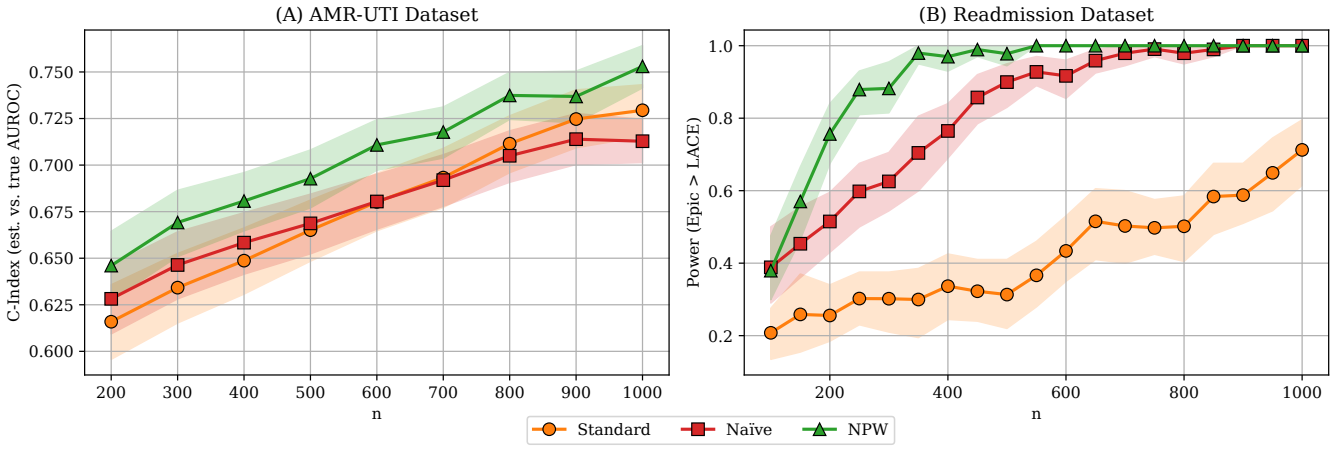

Figure 4: Empirical results with the real-world datasets. (A): C-index of model performance rankings induced by different AUROC estimates on the AMR-UTI dataset. (B): Statistical power for testing whether Epic outperforms LACE using different AUROC estimates on the Readmission dataset. On both figures, NPW achieves the highest C-index and statistical power across all sample size settings (n). Error bars are bootstrapped 95% confidence intervals.

shown in Figure 4(A), NPW augmented AUROC consistently improves over the standard approach when the sample sizes used to estimate nuisance parameters and estimate AUROC is less than or equal to 1000. As expected, increasing the number of samples diminishes the performance gap between NPW and standard AUROC. However, real-world RCT datasets often have limited sample sizes, underscoring the practical benefits of NPW in such scenarios.

Comparing Figures 2(B) and 4(A), we see that naïve approach improves model selection in the synthetic dataset, but worsens it in the AMR-UTI dataset. This is because the condition for naïve augmented AUROC to select the wrong model (Theorem 2) is more likely to hold in the AMR-UTI dataset, which supports our theoretical insights. See Appendix Section 6.2 for more details.

NPW Augmentation Boosts Statistical Power

Can NPW improve the statistical power of hypothesis testing in real-world RCTs? We use the Readmission RCT dataset to investigate this question. We aim to test whether the Epic model outperforms LACE in predicting readmission. Testing this hypothesis is challenging because the Epic model only moderately improves over LACE (Figure 3).

We evaluate the statistical power to test this hypothesis using different approaches and sample sizes. As shown in Figure 4(B), all methods yield higher power as the sample size increases, indicating more confidence in Epic's improvement over LACE. To achieve moderately high power (*i.e.*, 0.8), the standard approach requires over 1000 samples, of which 423 control samples are used, given the 58% treatment rate in the RCT. Under the same setting, the naïve approach, which uses all the data but is biased, requires close to 500 samples. In contrast, when using NPW, we obtain the same level of power with only 200 samples, a five-times improvement in sample efficiency over the standard approach.

Conclusion

In this work, we investigate model evaluation using AUROC with RCT data. We propose augmenting the standard evaluation protocol by leveraging treatment data to reduce estimation error and improve downstream model selection performance. We theoretically quantify the bias of naïve augmentation and derive the exact condition under which it leads to incorrect model selection. Leveraging these insights, we propose NPW, an unbiased approach that reweights data from the treatment group to mimic the distribution of individuals who would or would not experience the outcome in the absence of intervention. As a weighting-based approach, NPW generalizes to any binary classification metric. We empirically validate NPW and demonstrate that it enhances AUROC estimation, model selection, and hypothesis testing across diverse synthetic and real-world datasets.

Our work is not without limitations. First, we focus on the RCT setting, where interventions are assigned at random. Although restrictive, this is often necessary since interventions are frequently assigned deterministically. Second, implementing NPW requires estimating nuisance parameters. Therefore, our empirical results heavily depend on the quality of the nuisance parameter estimates, which can vary between applications. Encouragingly, our approach yields better estimates of model performance in a real-world setting with limited data. Practitioners applying NPW to their problem should first rigorously assess the quality of their nuisance parameter estimates. Finally, theoretical insights in this work focus on analyzing the bias associated with the estimates. Future work should examine questions regarding the variance of augmented AUROCs.

Despite these limitations, researchers evaluating models with RCTs should consider using NPW augmentation. This approach theoretically yields unbiased estimates of model performance and can make the most of RCT data, potentially reducing the length and cost of RCT studies.

Acknowledgements

This work was supported by the University of Michigan's AI & Digital Health Innovation, the Michigan Institute for Data & AI in Society's Propelling Original Data Science Grant, and the National Science Foundation (NSF) under Award No. 2124127. We thank Trenton Chang, Donald Lin, Donna Tjandra, Gregory Kondas, Jung Min Lee, Meera Krishnamoorthy, Michael Ito, Paco Haas, Sarah Jabbou, Stephanie Shepard, and Zhiyi Hu for their helpful conversations and feedback on this work.

References

Adnan, M.; Habib, A.; Ashraf, J.; Mussadiq, S.; Raza, A. A.; Abid, M.; Bashir, M.; and Khan, S. U. 2021. Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *Ieee Access*, 9: 7519–7539.

Amarasingham, R.; Patel, P. C.; Toto, K.; Nelson, L. L.; Swanson, T. S.; Moore, B. J.; Xie, B.; Zhang, S.; Alvarez, K. S.; Ma, Y.; et al. 2013. Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. *BMJ quality & safety*, 22(12): 998–1005.

Ayer, T.; Zhang, C.; Bonifonte, A.; Spaulding, A. C.; and Chhatwal, J. 2019. Prioritizing hepatitis C treatment in US prisons. *Operations Research*, 67(3): 853–873.

Boyer, C. B.; Dahabreh, I. J.; and Steingrimsson, J. A. 2023. Assessing model performance for counterfactual predictions. *arXiv preprint arXiv:2308.13026*.

Coston, A.; Mishler, A.; Kennedy, E. H.; and Chouldechova, A. 2020. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 582–593.

Deo, S.; Rajaram, K.; Rath, S.; Karmarkar, U. S.; and Goetz, M. B. 2015. Planning for HIV screening, testing, and care at the veterans health administration. *Operations research*, 63(2): 287–304.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Gerum, P. C. L.; Altay, A.; and Baykal-Gürsoy, M. 2019. Data-driven predictive maintenance scheduling policies for railways. *Transportation Research Part C: Emerging Technologies*, 107: 137–154.

Hansen, L. O.; Young, R. S.; Hinami, K.; Leung, A.; and Williams, M. V. 2011. Interventions to reduce 30-day rehospitalization: a systematic review. *Annals of internal medicine*, 155(8): 520–528.

Huang, Y.; Talwar, A.; Chatterjee, S.; and Aparasu, R. R. 2021. Application of machine learning in predicting hospital readmissions: a scoping review of the literature. *BMC medical research methodology*, 21: 1–14.

Hwang, A. B.; Schuepfer, G.; Pietrini, M.; and Boes, S. 2021. External validation of EPIC's Risk of Unplanned Readmission model, the LACE+ index and SQLape as predictors of unplanned hospital readmissions: A monocentric, retrospective, diagnostic cohort study in Switzerland. *PLoS One*, 16(11): e0258338.

Keogh, R. H.; and Van Geloven, N. 2024. Prediction under interventions: evaluation of counterfactual performance using longitudinal observational data. *Epidemiology*, 35(3): 329–339.

Ling, C. X.; Huang, J.; and Zhang, H. 2003. AUC: a better measure than accuracy in comparing learning algorithms. In *Advances in artificial intelligence: 16th conference of the Canadian society for computational studies of intelligence, AI 2003, halifax, Canada, June 11–13, 2003, proceedings 16*, 329–341. Springer.

Mac Iver, M. A.; Stein, M. L.; Davis, M. H.; Balfanz, R. W.; and Fox, J. H. 2019. An efficacy study of a ninth-grade early warning indicator intervention. *Journal of Research on Educational Effectiveness*, 12(3): 363–390.

Moore, T. J.; Heyward, J.; Anderson, G.; and Alexander, G. C. 2020. Variation in the estimated costs of pivotal clinical benefit trials supporting the US approval of new therapeutic agents, 2015–2017: a cross-sectional study. *BMJ open*, 10(6): e038863.

Oberst, M.; Boominathan, S.; Zhou, H.; Kanjilal, S.; and Sontag, D. 2020. AMR-UTI: antimicrobial resistance in urinary tract infections (version 1.0. 0).

Pajouheshnia, R.; Peelen, L. M.; Moons, K. G.; Reitsma, J. B.; and Groenwold, R. H. 2017. Accounting for treatment use when validating a prognostic model: a simulation study. *BMC medical research methodology*, 17: 1–12.

Teh, R.; and Janus, E. 2018. Identifying and targeting patients with predicted 30-day hospital readmissions using the revised LACE index score and early postdischarge intervention. *JBI Evidence Implementation*, 16(3): 174–181.

Van Walraven, C.; Dhalla, I. A.; Bell, C.; Etchells, E.; Stiell, I. G.; Zarnke, K.; Austin, P. C.; and Forster, A. J. 2010. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Cmaj*, 182(6): 551–557.

Wang, K.; Verma, S.; Mate, A.; Shah, S.; Taneja, A.; Madhiwalla, N.; Hegde, A.; and Tambe, M. 2023. Scalable decision-focused learning in restless multi-armed bandits with application to maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12138–12146.

Yang, Z.; Zeng, Z.; Wang, K.; Wong, S.-S.; Liang, W.; Zanin, M.; Liu, P.; Cao, X.; Gao, Z.; Mai, Z.; et al. 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of thoracic disease*, 12(3): 165.

Yeter, B.; Garbatov, Y.; and Soares, C. G. 2020. Risk-based maintenance planning of offshore wind turbine farms. *Reliability Engineering & System Safety*, 202: 107062.

Yu, S.; Farooq, F.; Van Esbroeck, A.; Fung, G.; Anand, V.; and Krishnapuram, B. 2015. Predicting readmission risk with institution-specific prediction models. *Artificial intelligence in medicine*, 65(2): 89–96.