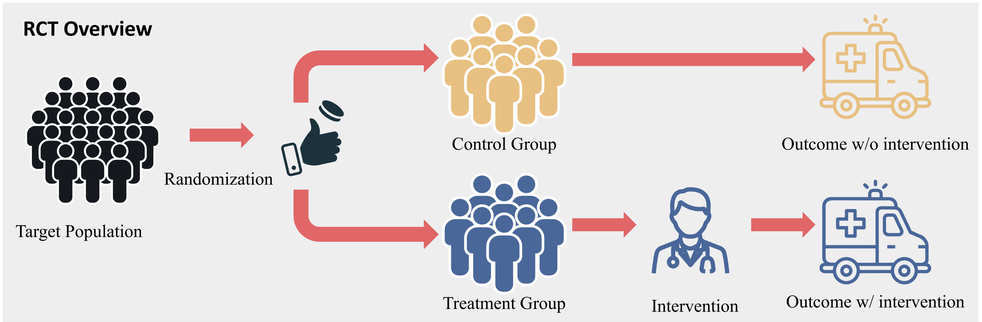
Measuring Model Performance in the Presence of an Intervention

Winston Chen[1], Michael Sjoding[2], Jenna Wiens[1] 1. University of Michigan Computer Science and Engineering, 2. University of Michigan Pulmonary and Critical Care

Motivation: Evaluating model's ability to predict outcome without intervention requires data from a randomized control trial (RCT), which is often expensive to conduct.



Problem Statement: How to estimate model's performance under no intervention with RCT data?

Task Inputs

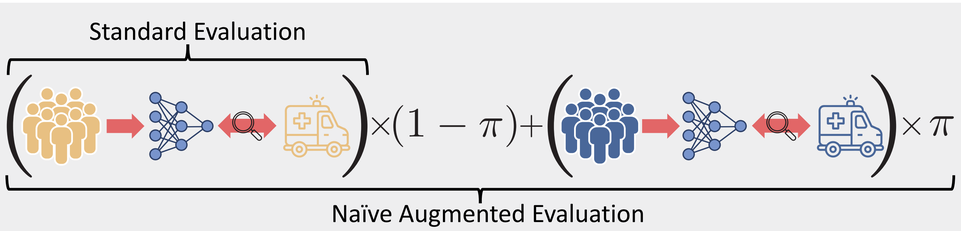
$x_i \sim \mathbf{P}(X)$ $t_i \sim \text{Bern}(\pi)$

$y_i(t_i) \sim \text{Bern}(\omega(X) + t_i\tau(X))$

Target Estimand

$$\mathbb{E}\left[\text{Performance metric}(\text{Model under evaluation})\right]$$

Gap: standard evaluation is unbiased but only uses data from the control group; naïvely augmenting it with data from the treatment group introduces bias.



Theorem 1: bias of naïve augmented AUROC.

When using AUROC as the metric, the bias of naïve augmented evaluation is:

$$\text{Bias}\left(\text{AUROC}_{\text{naïve}}(f)\right) = \alpha\delta(f) - \beta\sigma(f)$$

RCT specific parameters

Model under evaluation

Model's true AUROC

Model output's correlation with intervention's effect

Proposed Approach: Nuisance Parameter Weighting (NPW), a novel evaluation approach leveraging all RCT data to produce unbiased performance estimates via nuisance parameters.

Step 1. Estimate Nuisance Parameters

1. $\omega(X) = \mathbf{P}(Y|X, T=0)$ (W)

Outcome probability without intervention

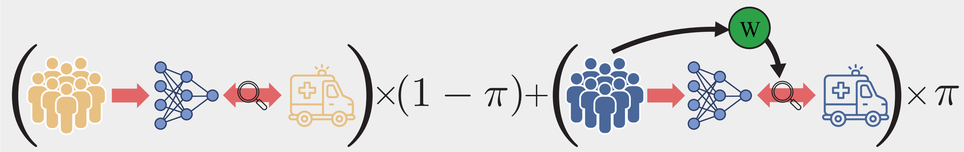
2. $\tau(X) = \mathbf{P}(Y|X, T=1) - \omega(X)$

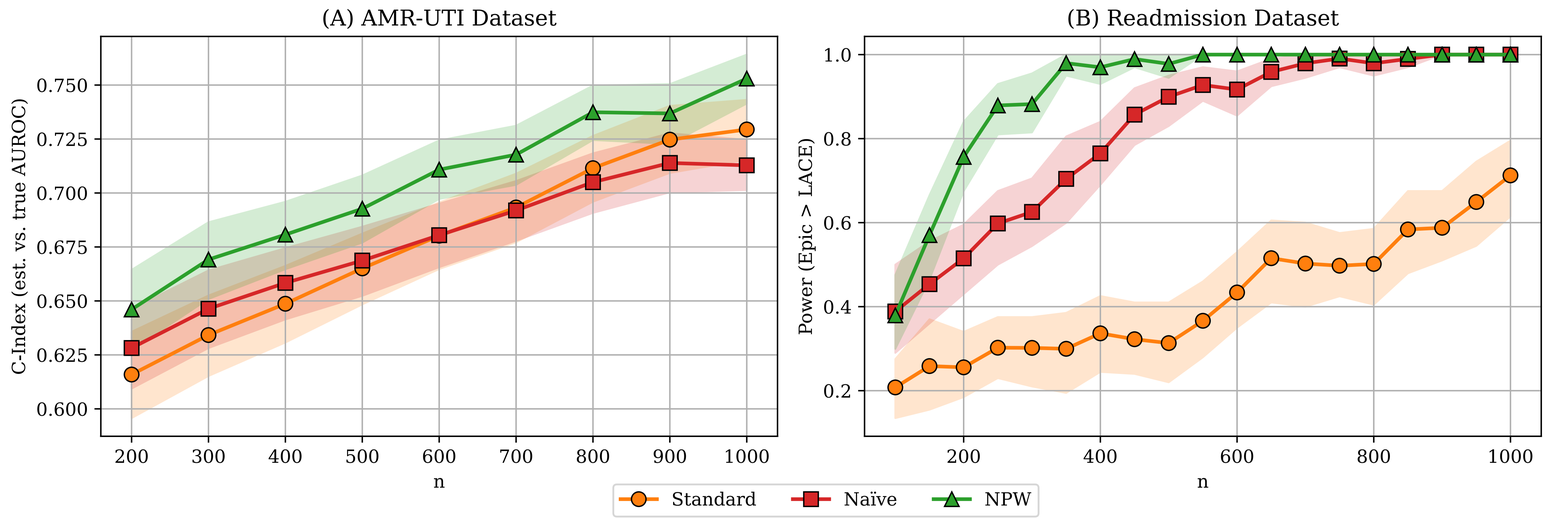
Conditional average treatment effect (CATE)

Note: Given RCT data, any supervised learning method guarantees unbiased estimates!

Step 2. Reweight the treatment data with nuisance parameter estimates

- *Given unbiased nuisance parameter estimates, NPW removes the evaluation bias from naïvely incorporating treatment data.*





Empirical Results: NPW improves real-world model evaluation:

- In the **AMR-UTI dataset [1]**, NPW produces more accurate model ranking, measured in C-index.
- In the **Michigan Medicine's Readmission dataset**, NPW achieves higher statistical power at differentiating the performance between two readmission prediction models (i.e., LACE & Epic).

Conclusion: researchers evaluating models with limited RCT data should consider using NPW to improve sample efficiency!

Reference:

[1] Oberst, M, et al. (2020). AMR-UTI: Antimicrobial Resistance in Urinary Tract Infections. *PhysioNet*.

Funded by University of Michigan's AI & Digital Health Innovation and MIDAS PODS Grant

chenwt@umich.edu