Evaluate Causal Models without Counterfactual on Real Data

Anonymous submission

Abstract

The evaluation of individual treatment effect (ITE) prediction is the most critical challenge for causal learning. The key issue is that only one potential outcome can be measured for an individual in reality, even for test data in randomized trials, since one cannot be treated and untreated simultaneously. The evaluation of existing ITE models mainly relies on synthetic outcomes and nearest neighbor matching. However, synthetic outcomes cannot be generalized to the real world well, and the evaluation of matched individuals is also not trivial. Therefore, it is desired that an evaluation scheme can directly judge which learned ITE model has a more accurate prediction for both individuals and population. In this paper, we propose comparing the ITE prediction error of different models on real data for both individuals and population. The ITE model input is pretreatment and treatment, and with the outcome as the model output, ITE can be calculated by difference. The outcome prediction error with the same treatment for factual and counterfactual individuals is independently distributed on the test dataset. To infer the distribution from factual data, we also assume that the factual indicators are independent of potential errors, which can be regarded as satisfied when the treatment is a simple randomization. Therefore, two models with given predicted potential outcomes and factual dataset can be evaluated by the confidence level that ITE prediction (mean) square error of one model is smaller than or equal to the other model for both individuals and population. Experimental studies show the effectiveness of the proposed evaluation scheme, and comparisons on a real dataset ALERT also show improving the model's performance in individual-level is more difficult than in population-level.

Introduction

In various real-world scenarios, there exist numerous applications that require the learning of a model to estimate individual treatment effect (ITE), such as personalized medicine, user conversion, recommendation system, marketing, and political elections. The most critical challenge of causal learning is to support the evaluation of research contributions for various causal models based on real scenarios (Cheng et al. 2022). In this paper, we focus on the fundamental problem of causal learning evaluation: "It is impossible to observe the value of $Y_t(u)$ and $Y_c(u)$ on the same unit, and therefore it is impossible to observe the effect of t on u" (Holland 1986). The fundamental problem exists even in randomized trial data. This raises an important question:

Given two causal models without the counterfactual outcome on real data, how can we infer which models' (mean) square prediction error of ITE is smaller for both the population and individuals?

To investigate this question, there are usually two kinds of model evaluation schemes in existing works: synthetic outcome and nearest neighbor matching. For example, (Hill 2011) uses linear/exponential Gaussian models as response surface A/B in Infant Health and Development Program (IHDP) dataset where models' parameters are randomly selected. The treatment is intensive high-quality child care and home visits, and the outcome is synthetic cognitive test scores. However, the real outcome function can not be abstracted from the real outcome by the evaluation on synthetic data. Another evaluation method is to use matching to impute counterfactual. The outcome of an individual's nearest neighbor in the opposite treatment group was used as a surrogate for the counterfactual outcome. However, the pair can be misspecified and counterfactual outcome of an individual may not be in the opposite treatment group sometimes. Also, the evaluation of nearest neighbor choosing, such as propensity score, is not trivial (Cheng et al. 2022). A specific example of matching is the twin data. For example, (Louizos et al. 2017) uses the mortality of one twin as the counterfactual mortality of the other twin, and birth weight is regarded as treatment. However, birth weight is not randomized so bias may exist. Also, the large quantity of treatment that need to be evaluated and the limited twins number in reality make the data availability very low. In order to make the learned causal models can be evaluated by real outcome so that they can be generalized to real scenes, we need to discover a new evaluation scheme with weaker assumptions and realistic data availability.

In this paper, we introduce a new causal model evaluation scheme which can help to learn and provide causal information from real outcome rather than from the function given by experts. The basic idea is that the outcome prediction errors with the same treatment (which we called potential error) for the model to be evaluated are independent identically distributed on the test dataset as shown in figure 1. The potential error distribution can be inferred from the observed sample if the treatment is randomized. We propose a Monte Carlo algorithm to calculate the confidence for arbitrary potential error distribution, such as histogram distribution. Fi-

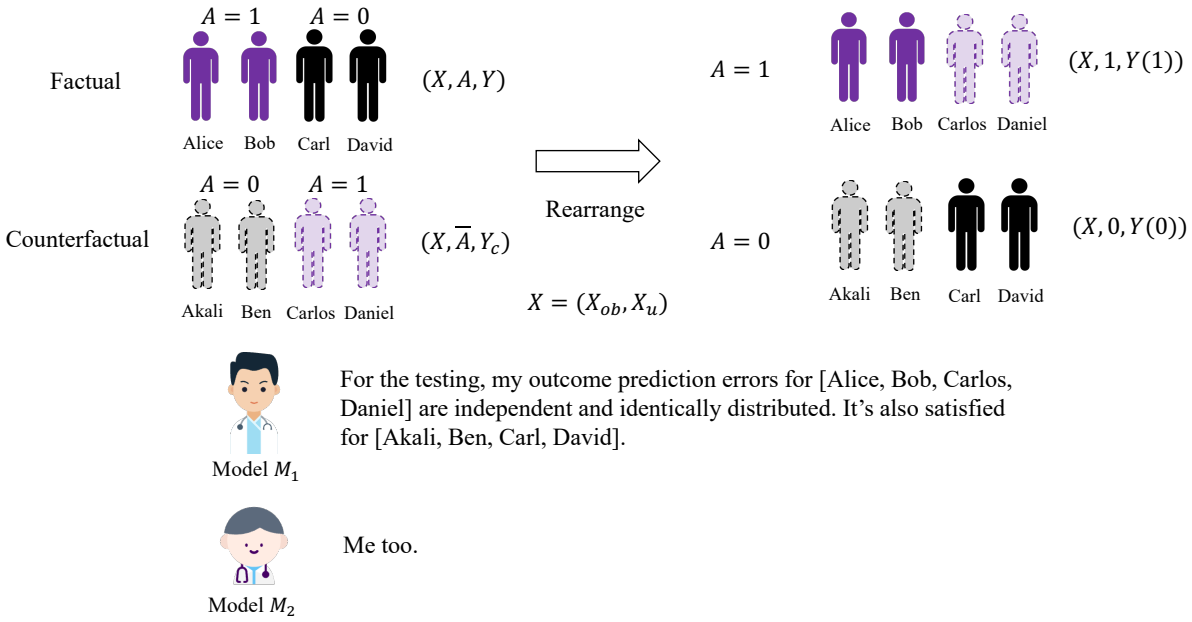


Figure 1: IID assumption of potential errors $\varepsilon_M^i(A)$. Potential errors follow $(\varepsilon_{M_1}(1), \varepsilon_{M_1}(0))$, $(\varepsilon_{M_2}(1), \varepsilon_{M_2}(0))$ respectively.

nally, the probability that one model's ITE (mean) square prediction error is smaller than or equal to the other can be calculated from real data for all individuals and populations.

Furthermore, we propose a metric that was called *popularity* of causal model based on the individuals' confidence difference of the two models, and we use average popularity on different classes in addressing the imbalanced outcome in the classification task.

The evaluation scheme bridges the gap between factual prediction and ITE prediction in causal learning. Our evaluation scheme does not require the models that need to be compared to have the same input format and counterfactual forms but it relies on the efficient inference of potential error distribution.

Our contribution can be concluded as follows,

- We propose a novel evaluation architecture for general causal models from both population-level and individual-level on *real* data where counterfactual is unknown. The potential errors with the same potential treatment are assumed as independently identical distributed conditional on the given model.
- We analyze the potential error with Gaussian distribution for continuous outcome and propose a Monte Carlo algorithm for arbitrary potential error distributions. The factual indicator is assumed to be independent from potential error to make the potential error distribution can be inferred directly from the factual errors. The assumption can be regarded as satisfied if the binary treatment is from simple randomization.
- We perform experiments for three classic models on the dataset ALERT with real outcomes. The experimental results show that CatBoost and Linear/Logistic models are significantly better than average/stochastic for ITE prediction in population-level and CatBoost performs better than others in individual-level although the popularity/prestige difference between CatBoost and Linear/Logistic is not significant.

Related works

ITE evaluation

The ITE evaluation approaches for causal models can be divided into two categories: individual matching, and synthetic outcome.

Individual matching is to find a factual outcome in an individual's opposite treatment group as the individual's counterfactual outcome. The key to the matching approach is the evaluation of the matching function. For example, BLR/BLN (Johansson, Shalit, and Sontag 2016) uses the outcome of the nearest neighbor of pretreatment to impute counterfactual data which is difficult to evaluate. A specific example is twin data. (Louizos et al. 2017) uses one twin's factual outcome as the other twin's counterfactual outcome despite of the low data availability.

Synthetic outcome is to assign a known function with random parameters which input is pretreatment and treatment, and output is the outcome. (Hill 2011) use two linear/logistic Gaussian models to generate potential outcomes by pretreatment and treatment. (Johansson, Shalit, and Sontag 2016) uses two linear Gaussian models to generate readers' experience from the hidden topic distribution Z of consumers' features X and the viewing device is generated by a softmax function from the hidden topic. However, the learned model's performance can not be generalized to real world well because the outcome function is what we want to learn from the real world data in causal learning. And if we know

well about the outcome, then the treatment effect can be induced directly by the given outcome function, there is no need to learn from the data.

There are also some other works of ITE evaluation for causal models. For example, uplift curve and Qini curve were often used for evaluation in uplift task (Gutierrez and Gérardy 2017). (Shalit, Johansson, and Sontag 2017) use the unemployment prediction performance in random trials as a metric of the causal model's performance. Recently, (Gentzel, Pruthi, and Jensen 2021) uses OSRCT data to predict individual-level outcomes under intervention but it can not be used to evaluate ITE prediction performance. Also, they did not reveal the connection between factual prediction and ITE prediction.

Causal models

The causal models are used to estimate the treatment effect, especially for individuals. For example, treatment agnostic representation network (TARNet) (Shalit, Johansson, and Sontag 2017) uses twin neural networks with different parameters to model the treated and control outcomes, respectively. CEVAE (Louizos et al. 2017) uses a non-parametric causal diagram prior to factorizing the causal effect into observation probability. CatBoost (Prokhorenkova et al. 2018) is one of state-of-the-art tree models for heterogeneous tabular dataset, which supports categorical and continuous features, especially for click rate prediction, and uplift modeling.

Notation and problem

Notation

For an individual i in the test dataset D_{te}, $X^i = (X^i_{ob}, X^i_u)$ is the pretreatment where X^i_{ob} is observed and X^i_u is unobserved, Y^i_{ob} is observed/measured/factual outcome, A^i is assignment of treatment, and Y_u is the unmeasured/counterfactual outcome. M is a model which input is treatment assignment A and features X, output is denoted as Y. The potential outcome and potential error conditional on model M for individual i are denoted as $Y^i_M(A)$ and $\varepsilon^i_M(A)$ where one of $Y^i_M(A)$ or $\varepsilon^i_M(A)$ is the factual outcome or factual error and others are counterfactual outcomes or counterfactual errors. And, X^i, X^i_{ob}, X^i_u, Y^i_{ob}, A^i, $Y^i_M(A)$, $\varepsilon^i_M(A)$ are all random variables. The treatment assignment A is binary in this paper.

Problem Description

Given two learned models $M_1(A, X_1)$ and $M_2(A, X_2)$ that need to be evaluated where features $X_1 \subseteq X$ and features $X_2 \subseteq X$, the target in individual level is to compare the square prediction error of ITE for M_1 and M_2, that is, whether $(M_1(1, X^i_1) - M_1(0, X^i_1) - (Y^i_{M_1}(1) - Y^i_{M_1}(0)))^2$ is smaller or equal to $(M_2(1, X^i_2) - M_2(0, X^i_2) - (Y^i_{M_2}(1) - Y^i_{M_2}(0)))^2$. The target in population level is to compare mean squared prediction error of ITE for M_1 and M_2, that is, whether $\frac{1}{n}\sum_{i=1}^{n}(M_1(1, X^i_1) - M_1(0, X^i_1) - (Y^i_{M_1}(1) - Y^i_{M_1}(0)))^2$ is smaller or equal to $\frac{1}{n}\sum_{i=1}^{n}(M_2(1, X^i_2) - M_2(0, X^i_2) - (Y^i_{M_2}(1) - Y^i_{M_2}(0)))^2$. The fundamental problem of the evaluation is that only one of the two potential outcomes $(Y^i_M(1), Y^i_M(0))$ can be measured even in a test dataset and randomized trial. Here, we assume that the measured potential outcome Y^i_{ob} of any individual i is independent from any given model M.

Methodology

Our evaluation scheme is as shown in figure 2. It is composed by two assumptions (gray), potential error inference (orange), and confidence calculation (yellow). We will first introduce the confidence calculation module, and then illustrate the potential error inference module. The ITE evaluation metric is introduced in the experiment section.

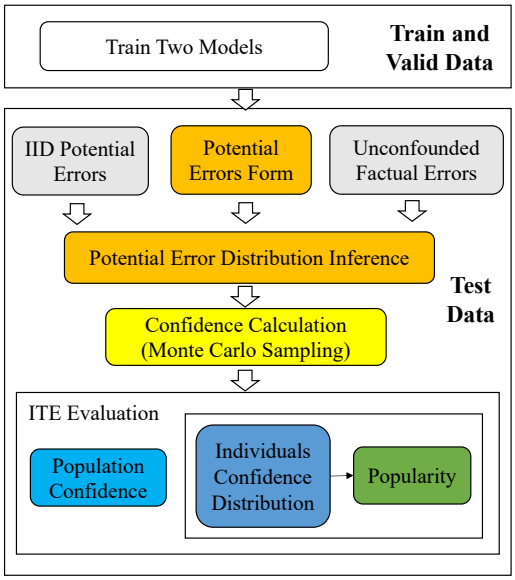


Figure 2: Proposed evaluation scheme.

Causal Model Evaluation by Potential Error Distribution

In this section, we demonstrate our scheme to deal with the fundamental problem in evaluation. The core idea of our solution is that the prediction error of potential outcome with the same treatment for any model to be evaluated is independent identically distributed on the test dataset as shown in figure 1.

For a model to be evaluated, if the potential error is not independently identical distributed on the test dataset, then some error patterns and structures that should be incorporated into the model may not be considered. Ignoring the presence of non-i.i.d. errors can lead to biased model estimates, and unreliable predictions.

Here is the formal statement of our i.i.d. potential error assumption,

Assumption 1 *For any model M to be evaluated, any individual i in the test dataset D_{te}, and any potential treatment a in the range of A^i for individual i, we have* $\varepsilon^i_M(a) \overset{i.i.d.}{\sim}$

$\mathcal{F}(a, M, D_{te})$, where $\mathcal{F}$ represents the underlying probability distribution function governing the outcome prediction errors which is determined by potential treatment a, model to be evaluated M, and population D_{te} in the test dataset.

The i.i.d. potential error assumption gives the generalization condition of evaluation results based on the given models and real data in the test dataset. The inference of the distribution $\mathcal{F}$ from factual dataset $D_{te} = \{A^i_{ob}, X^i_{ob}, Y^i_{ob}\}^i$ will be introduced in the next subsection where A^i_{ob} is the factual treatment for individual i. We introduce our approach to evaluate the individual-level error of ITE and population-level of ITE in this subsection.

Evaluate Causal Models for Individual For an individual i, the question of interest is whether the absolute error of ITE prediction for model M_1 is lower than or equal to M_2 for an individual i. Given the potential error distributions $(\varepsilon_{M_1}(A^i_u), \varepsilon_{M_1}(A^i_u))$, model pair (M_1, M_2), and factual dataset D_{te} where A^i_u is counterfactual treatment for the individual i, the probability can be calculated absolutely.

Here, theorem 1 demonstrates the specific case when the potential error distributions ($\varepsilon_{M_1}(A^i_u)$,$\varepsilon_{M_2}(A^i_u)$) are both Gaussian distributions.

Theorem 1 *Without loss of generality, let the factual treatment $A^i_{ob} = 1$ for an individual i. Given the individual's pretreatment X^i_{ob} and its factual outcome $Y^i_{ob} = Y^i(1)$, and given two models $Y = M_1(A, X_1)$ and $Y = M_2(A, X_2)$ to be evaluated, the probability that square ITE prediction error of model M_1 is smaller than or equal to model M_2, can be represented as a cumulative distribution function of a generalized non-central chi-squared distribution $F_x(x = 0; \tilde{\chi}^2(\boldsymbol{w}, \boldsymbol{k}, \boldsymbol{\lambda}, 0, 0))$ where $\boldsymbol{w} = [\sigma_1^2, -\sigma_2^2]$, $\boldsymbol{k} = [1, 1]$, and $\boldsymbol{\lambda} = \left[(\frac{\epsilon^i_{M_1}(1)-\mu_1}{\sigma_1})^2, (\frac{\epsilon^i_{M_2}(1)-\mu_2}{\sigma_2})^2\right]$ if potential error distributions $\varepsilon^i_{M_1}(0)$ and $\varepsilon^i_{M_2}(0)$ follow Gaussian distribution $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, and $\epsilon^i_{M_1}(1)$, $\epsilon^i_{M_2}(1)$ are factual prediction errors.*

Evaluate Causal Models for Population For a population D_{te}, the question of interest is whether the mean square error of the ITE prediction for model M_1 is lower than or equal to the model M_2. Given the potential error distributions $(\varepsilon_{M_1}(A^i_u), \varepsilon_{M_1}(A^i_u))$ for all individuals i, model pair (M_1, M_2), and factual dataset D_{te} where A^i_u is the counterfactual treatment for individual i, the probability is absolutely can be calculated.

Here, theorem 2 demonstrates the specific case when the potential error distributions for all individuals are all Gaussian distributions.

Theorem 2 *Given a factual test dataset $D_{te} = (X^i_{ob}, A^i_{ob}, Y^i_{ob})^i$, and given two models $Y = M_1(A, X_1)$ and $Y = M_2(A, X_2)$ to be evaluated, the probability that mean square ITE prediction error of the model M_1 is smaller than or equal to the model M_2 can be represented as a cumulative distribution function of a generalized non-central chi-squared distribution $F_x(x = 0; \tilde{\chi}^2(\boldsymbol{w}, \boldsymbol{k}, \boldsymbol{\lambda}, 0, 0))$ where $\boldsymbol{w} = \left[\underbrace{\sigma^2_{10}..\sigma^2_{10}}_{n_1}, \underbrace{\sigma^2_{11}..\sigma^2_{11}}_{n_0}, \underbrace{-\sigma^2_{20}.. - \sigma^2_{20}}_{n_1}, \underbrace{-\sigma^2_{21}.. - \sigma^2_{21}}_{n_0}\right]_{1*2n}$, $\boldsymbol{k} = [1, ...1]_{1*2n}$, $\boldsymbol{\lambda} = [\boldsymbol{\lambda_{11}}, \boldsymbol{\lambda_{10}}, \boldsymbol{\lambda_{21}}, \boldsymbol{\lambda_{20}}]_{1*2n}$, and $\boldsymbol{\lambda_{11}} = \left[(\frac{u_{10}-\epsilon^1_{M_1}(1)}{\sigma_{10}})^2..(\frac{u_{10}-\epsilon^{n_1}_{M_1}(1)}{\sigma_{10}})^2\right]_{1*n_1}$, $\boldsymbol{\lambda_{10}} = \left[(\frac{u_{11}-\epsilon^{n_1+1}_{M_2}(0)}{\sigma_{11}})^2..(\frac{u_{11}-\epsilon^{n}_{M_2}(0)}{\sigma_{11}})^2\right]_{1*n_0}$, $\boldsymbol{\lambda_{21}} = \left[(\frac{u_{20}-\epsilon^1_{M_2}(1)}{\sigma_{20}})^2..(\frac{u_{20}-\epsilon^{n_1}_{M_2}(1)}{\sigma_{20}})^2\right]_{1*n_1}$, $\boldsymbol{\lambda_{20}} = \left[(\frac{u_{21}-\epsilon^{n_1+1}_{M_2}(0)}{\sigma_{21}})^2..(\frac{u_{21}-\epsilon^{n}_{M_2}(0)}{\sigma_{21}})^2\right]_{1*n_0}$ if potential error distribution $\varepsilon_{M_1}(1)$, $\varepsilon_{M_2}(1)$, $\varepsilon_{M_1}(0)$, $\varepsilon_{M_2}(0)$ follow Gaussian distribution $N(\mu_{11}, \sigma_{11})$, $N(\mu_{21}, \sigma_{21})$, $N(\mu_{10}, \sigma_{10})$, $N(\mu_{20}, \sigma_{20})$, and $\epsilon^i_{M_1}(A^i_{ob})$, $\epsilon^i_{M_1}(A^i_{ob})$ are factual prediction errors.*

The numerical solution of the cumulative distribution function (CDF) for generalized non-central chi-square distribution (Davies 1980) (Davies 1973) can be calculated in python (Danilo Horta 2020).

Evaluate Casual model with Arbitrary Potential Error Distributions In addressing arbitrary potential error distributions, we propose a Monte Carlo algorithm to calculate the individuals' confidences $\mathbf{C}^j$ that square error of ITE prediction for model M_1 is smaller or equal to model M_2, and population confidence p that mean square error of ITE prediction model M_1 is smaller or equal to model M_2 as showed in algorithm 1.

Evaluate Causal Models for Outcome Classification For binary classification, the potential error distribution of outcome is not a continuous random variable but a discrete random variable with range $[-1, 0, 1]$. The distance between the prediction score and factual prediction is not included. Our evaluation for outcome classification is based on the potential error distribution of the probability score. The histogram distribution with range $[-1, 1]$ is applied in our evaluation for the classification task.

Infer Potential Error Distribution from Factual Data

If potential error distributions $\varepsilon^i_M(A)$ were given, the confidence of model evaluation can be calculated by the approach in the last subsection. However, we only have an observed sample of the potential error distribution due to unmeasured counterfactual error which is introduced by the fundamental problem of causal learning. The question is how to infer the potential error distribution from the observed sample.

First, we demonstrate the measurement process of the observed sample from the population by a factual indicator $\mathbb{I}$,

Definition 1 ***Factual Indicator*** $\mathbb{I}$. *For a given model $Y = M(A, X)$ and a given population $D_{te}(A)$ with mixed factual and counterfactual individuals for all potential treatments, a factual indicator $\mathbb{I}^i_M(A)$ is a random variable whose input is an individual i with potential error $\epsilon^i_M(A)$ and output is a binary value to represent whether the potential error $\epsilon^i_M(A)$ of this individual i is a factual error.*

Algorithm 1: Proposed Monte Carlo algorithm to calculate individual and population confidence for models' evaluation

input : potential error distributions $\varepsilon_{M_1}(1)$, $\varepsilon_{M_1}(0)$, $\varepsilon_{M_2}(1)$, $\varepsilon_{M_2}(0)$, factual errors $\boldsymbol{\epsilon^{i}_{M_1}(\mathbf{A^{i}_{ob}})}$, $\boldsymbol{\epsilon^{i}_{M_2}(\mathbf{A^{i}_{ob}})}$, individual number N, sampling number M, treatment number $K = 2$, factual indicator $\mathbb{I}$

output: Individual confidences $\mathbf{C}$, population confidence p

$n_p = 0$
$\mathbf{n^j_c} \leftarrow \mathbf{0}$
for $i \leftarrow 1$ **to** M **do**
  $p_t = 0$
  for $j \leftarrow 1$ **to** N **do**
    $n_c = 0$
    for $k \leftarrow 1$ **to** K **do**
      if $\mathbb{I}^i(k) = 0$ **then**
        $\epsilon^i_{M_1}(k) \sim \varepsilon_{M_1}(k)$
      end
    end
    $t \leftarrow (\epsilon^i_{M_1}(1) - \epsilon^i_{M_1}(0))^2 - (\epsilon^i_{M_2}(1) - \epsilon^i_{M_2}(0))^2$
    $p_t \leftarrow p_t + t$
    if $t \leq 0$ **then**
      $n^j_c += 1$
    end
  end
  if $p_t \leq 0$ **then**
    $n_p += 1$
  end
end
$p \leftarrow \frac{n_p}{M}$
for $j \leftarrow 1$ **to** N **do**
  $\mathbf{C^j} \leftarrow \frac{\mathbf{n^j_c}}{M}$
end

We discuss the potential errors' inference in the test dataset with randomized treatment and non-randomized treatment separately.

Randomized treatment In a classical randomized experiment, the treatment assignment is a known function $Pr(A)$ which is probabilistic, individualistic, and uncounfounded (Imbens and Rubin 2015). We let the treatment assignment be a simple randomization $Pr(A) = 0.5$ for binary treatment without loss of generality.

Similar to uncounfounded assumption which assumes treatment assignment A^i is independent from potential outcomes $(Y^i(1), Y^i(0))$, we assume that factual indicator $\mathbb{I}^i_M(A)$ is independent from potential errors $\varepsilon^i_M(A)$ for any potential value of A,

Assumption 2 *For any individual i (including factual and counterfactual individual) in population $D_{te}(a)$ and any model M to be evaluated, we have $\mathbb{I}^i_M(a) \perp \varepsilon^i_M(a)$ for any $a \in Range(A)$.*

We should ensure the assumption is satisfied before the test dataset was collected. Here is an illustration of assumption 2 when observed treatment A^i_{ob} is randomized. Let $I^i_M(1) = A^i_{ob}$ and $I^i_M(0) = 1 - A^i_{ob}$, because the A^i_{ob} is randomized, and the model M is fixed before the testing, and $I^i_M(A)$ is a one-to-one function of A, so the $I^i_M(1)$ and $I^i_M(0)$ can also be regraded as randomized. If assumption 2 holds, then we can infer the potential error distributions from the observed sample directly if the treatment is randomized as the following theorem,

Theorem 3 *For a given model M to be evaluated, and any individual i, if $I^i_M(a) \perp \varepsilon^i_M(a)$ for any $a \in Range(A)$, then $Pr(\varepsilon^i_M(a)) = Pr(\varepsilon^i_M(a) | I^i_M(a) = 1)$.*

For example, given a random sample $[1, 0, 0, 1]$ of treatment assignment $Pr(A)$ from the population, it is also a random sample $[1, 0, 0, 1]$ of factual indicator $\mathbb{I}^i_M(1)$ and a random sample $[0, 1, 1, 0]$ of factual indicator $\mathbb{I}^i_M(0)$. According to theorem 3, we can infer the potential error distribution $\varepsilon^i_M(1) = N(\mu_1, \sigma_1)$ and $\varepsilon^i_M(0) = N(\mu_0, \sigma_0)$ from the observed errors where $\mu_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} M(1, X^i) - Y^i_M(1)$, $\sigma_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(M(1, X^i) - Y^i_M(1) - \mu_1)^2$, and $\mu_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} M(0, X^i) - Y^i_M(0)$, $\sigma_0 = \frac{1}{n_0 - 1}\sum_{i=1}^{n_0}(M(0, X^i) - Y^i_M(0) - \mu_0)^2$ where n_0 and n_1 are the number of factual individuals with $A^i = 0$ and $A^i = 1$ respectively. For histogram distributions, we use frequency with $\max(\sqrt{n_{te}}, 100)$ bins with range $[-1, 1]$ to infer them.

Non-randomized treatment In observation study, treatment assignment $Pr(A)$ is not known (Imbens and Rubin 2015), so the factual indicator $\mathbb{I}^i_M(A)$ is difficult to choose to satisfy the assumption 2. So, we can not directly infer the potential error from the observed error. (Rosenbaum and Rubin 2023) illustrates some principles for observation study design.

Experiment

In this section, we describe the evaluation result of our scheme for some existing models' ITE prediction performance on a randomized dataset ALERT with real outcomes. We use the confidence table of model matches to demonstrate the ITE prediction performance of matched models for the population. And we use individuals' confidence distribution about the model matches to visualize the ITE prediction performance of matched models for individuals.

Dataset

We use a heterogeneous tabular dataset ALERT for causal models' evaluation, which is introduced in (Gentzel, Pruthi, and Jensen 2021) for observational causal inference. The reason to use the ALERT is that it is derived from a double-blinded, multi-center, and parallel designed randomized trial that recorded electronic health record (EHR) data of patients (Wilson et al. 2021). The trial used simple randomization to evaluate the impact of an acute kidney injury (AKI) alert caused by the electronic system compared to usual care without an alert. So the assumption 2 can be reasonable.

The dataset comprises a total of 6,030 adult inpatients with AKI. It includes 49 pretreatments, 1 treatment, and 42 posttreatments. Among the 49 pretreatments, there are 28 discrete features and 23 continuous features. Within the cohort of 6,030 patients, 948 patients experienced AKI progression within a 14-day period, while 5,082 patients did not exhibit such progression.

In the prepossessing, we remove the individuals whose pretreatment includes at least a NAN value. After preprocessing, there are 5347 patients in which 4470 AKI progression after the alert and 877 did not exhibit such progression.

The interested task in the ALERT dataset is to predict the individual treatment effect of alert based on the available pretreatment variables. For the classification task, we choose AKI progression within 14 days as the outcome. For the regression task, we choose minimum systolic after 24 hours of alert as the outcome.

Implementation Details

To quantify the influence of dataset splitting randomness on models, all experiments were conducted $\sqrt{n}$ times using different random splits of the dataset where n is the number of individuals. Train/test splitting ratio is 8:2.

We evaluated three models for the regression task: average, linear model and CatBoostRegression (Prokhorenkova et al. 2018), and we evaluated three models for the classification task: stochastic, logistic regression, and CatBoostClassifier.

Regression: ITE from alert to minimum systolic

Model Learning and Potential Error Distribution Inference The model learning details for both the regression model and the classification model can be seen in the appendix. For continuous outcome, we use Q-Q (quantile-quantile) plot to check the Gaussian assumption of potential error, and we did not check the histogram assumption for discrete outcome. The detail of Q-Q plot can also be seen in the appendix.

Confidence Visualization In order to evaluate the models in both individual level and population level, we use different metrics.

- **Population Level.** Table 1 shows the population confidence of matched models for regression and classification, respectively. We can see that the linear and CatBoost are significantly more accurate than the average prediction. But the advantage of CatBoost compared with linear is not significant.
- **Individual Level.** In reality, the prediction MSE of ITE in the population level is usually not sufficient for trustworthy models. Figure 3 shows the individuals' confidence distribution for matched models.

  We create a metric ρ based on the individuals' confidence distribution to measure the popularity/prestige of the model M_1 comparing with a reference model M_0 which can be calculated by the following formula,

$$\rho = \frac{2\mathbf{card}(\mathbf{p}(|\mathbf{e}^i_{M_1}| \leq |\mathbf{e}^i_{M_0}|) > 0.5)}{n} - 1 \in [-1, 1] \tag{1}$$

  where n is individual number and ρ is the normalized difference between vote number of model M_1 and the reference model M_0. Table 2 shows the popularity among the three models. Linear model and CatBoost are significantly more accurate than the average model in the population-level. In the individual-level, the incremental individual vote is only about 75% with an average model as a reference. The difference between the linear model and CatBoost is also not significant at individual-level. Figure 3 shows the detailed individuals' confidence distributions.

$p(MSE_r \leq MSE_b)$	Linear	CatBoost
Average	1.0	1.0
Linear	1	$.594 \pm .135$

Table 1: Population confidence when the outcome is minimum systolic. r: right; b: bottom.

$\rho(M_r; M_b)$	Linear	CatBoost
Average	$.751 \pm .024$	$.749 \pm .029$
Linear	0	$.080 \pm .113$

Table 2: Popularity among matched models at individual-level when outcome is minimum systolic. r: right; b: bottom.

Classification: ITE from Alert to AKI progress

Confidence Visualization The histogram distribution does not have good forms as Gaussian. We use algorithm 1 to generate 1k random samples for all individuals in the test dataset to calculate the population-level confidence and individual-level confidence of matched models.

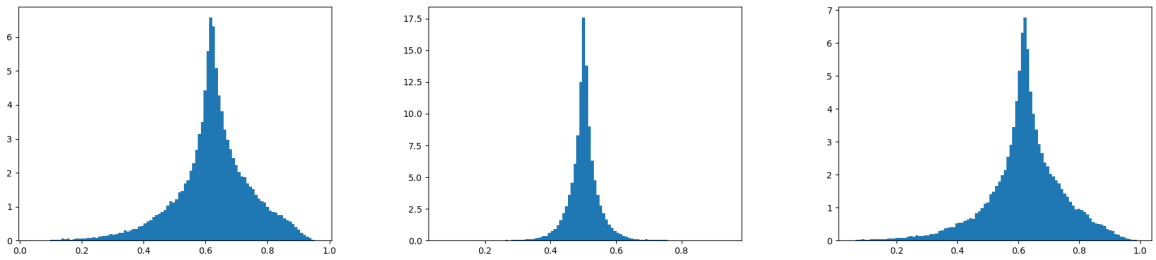
- **Population Level**. In addressing the imbalanced classes, the ITE difference between prediction and sampled real value is re-weighted by individual's classes, the line 13 in algorithm 1 was replaced by the following,

$$p_t \leftarrow p_t + \frac{t}{n_{Y=y^j}} \tag{2}$$

  Table 3 shows the balanced population confidence of matched models. The Logistic and CatBoost are significantly more accurate than stochastic, and CatBoost is significantly more accurate than the Logistic.
- **Individual Level**. In addressing the imbalanced classes, we use average popularity for different classes of the outcome,

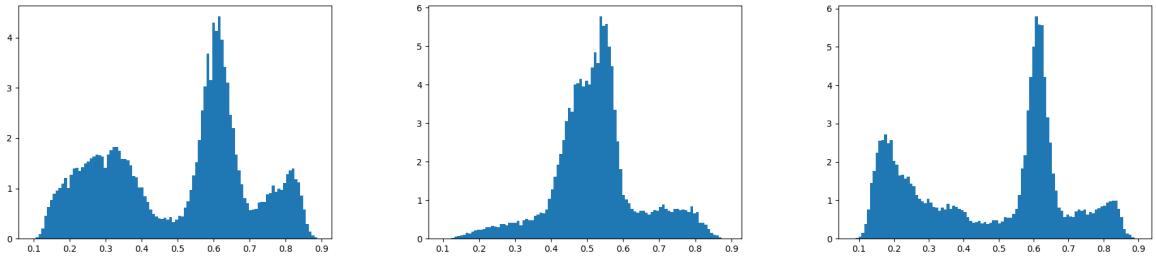
$$\rho = \frac{1}{K}\sum_{c=1}^{K} \rho_c \in [-1, 1] \tag{3}$$

  where K is the class number of the outcome and ρ_c is the popularity for individuals whose $Y_{ob} = c$. Table 4 shows the balanced popularity among the three models. The incremental individual votes of CatBoost comparing Stochastic is statistically significantly increased by a 5% level of confidence. Figure 4 shows the detailed individual confidence distribution for matched models.



(a) Catboost vs Average (b) Catboost vs Linear (c) Linear vs Average

Figure 3: Individual confidence distribution $p(|\epsilon^i_{M_1}| < |\epsilon^i_{M_0}|)$ for regression task. When p is larger than 0.5, it means the left model is more accurate than the right model for this individual. The x-axis is individual confidence, the y-axis is the individual density.



(a) Catboost vs Stochastic (b) Catboost vs Logistic (c) Logistic vs Stochastic

Figure 4: Individual confidence distribution $p(|e^i_{M_1}| \leq |e^i_{M_0}|)$ for classification task. When p is larger than 0.5, it means the left model is more accurate than the right model for this individual. The x-axis is individual confidence, the y-axis is the individual density. The individual number has been re-weighted by inverse of class numbers.

$p(MSE_r \leq MSE_b)$	Logistic	CatBoost
Stochastic	$.998 \pm .003$	$.999 \pm .001$
Logistic	1	$.924 \pm .060$

Table 3: Population confidence when the outcome is AKI progression in 14 days. r: right; b: bottom.

$\rho(M_r; M_b)$	Logistic	CatBoost
Stochastic	$.025 \pm .009$	$.032 \pm .008$
Logistic	0	$.039 \pm .025$

Table 4: Popularity among matched models at the individual level when the outcome is AKI progression in 14 days. r: right; b: bottom.

Conclusion

In this paper, we introduced a novel evaluation scheme of ITE prediction for general causal models based on real outcome. It can judge which model learns more causal information from the real outcome. Our framework is based on two assumptions: potential errors are independently identically distributed, and factual indicators are independent from the potential errors. We analyze potential error with Gaussian and propose a Monte Carlo method for arbitrary distributions. We performed experiments on a real dataset ALERT for some existing models from both individual-level and population-level. Our work bridges factual prediction and ITE prediction.

There are also some limits of our work. First, observed errors with the same treatment can be seen as a random sample when the treatment assignment is from simple randomization. However, it may not be in observation data so we can not infer the error distribution from observed errors directly. Second, although the Monte Carlo method can be used for arbitrary potential error distributions, more efficient algorithms are needed to accelerate the computation of matched causal model's confidence score. Third, for a new individual that is a uniform sampling from the same superpopulation but not in the test dataset, our evaluation always gives the same confidence for two models. A more individualized confidence calculation approach is needed, such as nearest neighbor matching of predicted ITE.

References

Cheng, L.; Guo, R.; Moraffah, R.; Sheth, P.; Candan, K. S.; and Liu, H. 2022. Evaluation methods and measures for

causal learning algorithms. *IEEE Transactions on Artificial Intelligence*, 3(6): 924–943.

Danilo Horta. 2020. chi2comb. https://github.com/limix/chi2comb-py. Accessed: 2023-08-01.

Davies, R. B. 1973. Numerical inversion of a characteristic function. *Biometrika*, 60(2): 415–417.

Davies, R. B. 1980. The distribution of a linear combination of $\chi 2$ random variables. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 29(3): 323–333.

Gentzel, A. M.; Pruthi, P.; and Jensen, D. 2021. How and why to use experimental data to evaluate methods for observational causal inference. In *International Conference on Machine Learning*, 3660–3671. PMLR.

Gutierrez, P.; and Gérardy, J.-Y. 2017. Causal inference and uplift modelling: A review of the literature. In *International conference on predictive applications and APIs*, 1–13. PMLR.

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

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396): 945–960.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Johansson, F.; Shalit, U.; and Sontag, D. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*, 3020–3029. PMLR.

Louizos, C.; Shalit, U.; Mooij, J.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6449–6459.

Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Rosenbaum, P.; and Rubin, D. 2023. Propensity scores in the design of observational studies for causal effects. *Biometrika*, 110(1): 1–13.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, 3076–3085. PMLR.

Wilson, F. P.; Martin, M.; Yamamoto, Y.; Partridge, C.; Moreira, E.; Arora, T.; Biswas, A.; Feldman, H.; Garg, A. X.; Greenberg, J. H.; et al. 2021. Electronic health record alerts for acute kidney injury: multicenter, randomized clinical trial. *Bmj*, 372.