Personalizing Treatment for Intensive Care Unit Patients with Acute Respiratory Distress Syndrome

Comparing the S-, T-, and X-learner to Estimate the Conditional Average Treatment Effect for High versus Low Positive End-Expiratory Pressure in Mechanical Ventilation

Juul Schnitzler[2]

Supervisors: Jesse Krijthe[2]**, Rickard Karlsson**[2]**, Jim Smit**[1,2]

[1]**Department of Intensive Care, Erasmus Medical Center, Rotterdam, The Netherlands**
[2]**EEMCS, Delft University of Technology, Delft, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Juul Schnitzler
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Rickard Karlsson, Jim Smit, Jasmijn Baaijens

Abstract

Mechanical ventilation is a vital supportive measure for patients with acute respiratory distress syndrome (ARDS) in the intensive care unit. An important setting in the ventilator is the positive end-expiratory pressure (PEEP), which can reduce lung stress but may also cause harmful side effects. This research investigates the personalization of PEEP settings based on patient characteristics using three meta-learning algorithms (S-, T-, and X-learner) to estimate the conditional average treatment effect. Additionally, the hypothesis that the X-learner performs particularly well under a significant imbalance in patient numbers between treatment groups is tested.

Results show that the X-learner slightly outperforms the S- and T-learners in terms of mean squared error under various unbalanced conditions in simulated data. However, the overall ability of these meta-learners to identify patients benefiting from high PEEP remains inconclusive. When using gradient boosted trees or random forest as base models, cumulative gain curves on MIMIC-IV data indicate potential overfitting. While the X-learner performs somewhat better on this data, the low area under the curve scores suggests a minimal distinction between high and low PEEP groups. External validation with data from a randomized control confirms that the models do not effectively distinguish between treatment groups. These findings suggest that further investigation with more complex models and real-world data is needed to validate the potential of meta-learning algorithms in personalizing PEEP settings for ARDS patients.

1 Introduction

In the intensive care unit (ICU), mechanical ventilation is a vital supportive measure for patients suffering from acute respiratory distress syndrome (ARDS) [1]. An important setting of the mechanical ventilator is the positive end-expiratory pressure (PEEP). High PEEP has the potential to mitigate lung stress and strain. However, these benefits may be negated by adverse side effects. This complicates the decision on how to set the PEEP in mechanical ventilation (high or low). Instead of using the same PEEP regime for all patients, some patients might benefit more from a certain PEEP regime than others. Accurately estimating the individual treatment effect of high versus low PEEP allows for the personalization of treatment based on patient characteristics, optimizing treatment effectiveness and improving patient survival outcomes.

Previous research regarding high versus low PEEP for ICU patients with ARDS focused on deciding whether to use a high or low PEEP regime in general. However, results were inconclusive [2–5]. Rather than a one-size-fits-all solution, it is hypothesized that some patient subgroups benefit more from high PEEP treatment. Findings from an analysis by Calfee et al. [6] showed the existence of two distinguishable sub-phenotypes of ARDS patients. More importantly, their findings have indicated that these two subgroups respond differently to low versus high PEEP. Findings from a meta-analysis by Briel et al. [7] comparing high versus low PEEP for patients with acute lung injury or ARDS, suggested that a high PEEP treatment was associated with a higher survival rate among the subgroup of patients with ARDS. Additionally, they showed that patients with less severe lung injury may experience harmful effects from high PEEP.

Given the findings discussed above, the current study aims to test the hypothesis that some patients benefit more from high PEEP compared to others. We compare different machine learning algorithms to examine whether the PEEP treatment can be personalized based on certain patient characteristics. These algorithms include three meta-algorithms: the S-learner, the T-learner, and the X-learner [8], with a primary focus on the latter. The performance of these algorithms will be compared in estimating the conditional average treatment effect (CATE) on simulated data, the MIMIC-IV dataset [9], and data from a randomized trial. The simulations aim to generate data that approximates real-life medical data to evaluate the performance of the S-, T-, and X-learner in various settings. The MIMIC-IV dataset provides data from an observational study on ICU patients with ARDS. The randomized trial data will be used for external validation. Secondly, it is hypothesized that the X-learner is particularly effective when the number of patients in the control group and the treatment group significantly differ [8]. This research additionally aims to test this hypothesis. Therefore, the main research question and sub-question are defined as follows:

Main Research Question:
How do the S-learner, T-learner, and X-learner perform in estimating the CATE to predict which ICU patients suffering from ARDS benefit from high PEEP compared to low PEEP in mechanical ventilation, based on patient characteristics?

Sub-Question:
Does the X-learner perform particularly well in estimating the CATE when the treatment assignment in the data is significantly unbalanced?

Section 2 of this paper provides a formal description of the problem and gives a definition and further explanation of the MIMIC-IV dataset, causal inference, confounding, and CATE. In section 3 we describe the conducted methodology, explain how the meta-learners work, and provide the set of confounders that were identified. Section 4 elaborates on the experimental setup by describing the experiments that were conducted and their results. In section 5 we discuss the meaning of these results, and in section 6 we reflect on the ethical aspects of this study. Finally, in section 7 we provide our conclusions and formulate recommendations for further research.

2 Problem Setup and Definitions

2.1 Problem description

The problem that is addressed in the current study is the heterogeneity of treatment effects in mechanical ventilation for ICU patients with ARDS. This research aims to determine whether the treatment strategy can be personalized based on

patient characteristics using the MIMIC-IV dataset by applying meta-learners to estimate the CATE and evaluate the impact of high versus low PEEP on patient survival outcomes.

2.2 MIMIC-IV dataset

The MIMIC-IV dataset is a publicly available ICU database [9], which includes observational data from a total of 3,941 patients suffering from ARDS. The treatment labels in the dataset are based on the PEEP regimes (the FiO2/PEEP tables) defined in the randomized trial by Brower et al [10]. Patients were classified according to the FiO2/PEEP combinations observed throughout their entire ICU stay. Specifically, 12% of the patients were labeled under a 'high' PEEP regime, with the remaining 88% labeled under a 'low' PEEP regime. There are several covariates available for each patient in the database, including the PEEP regime, the 28-day mortality, demographic data (e.g., age, sex, weight), and medical data (e.g., heart rate, lung compliance).

2.3 Causal inference

To test our main hypothesis, we examine the difference in 28-day mortality (the outcome) from following a high PEEP versus following a low PEEP (the treatment). The treatment variable will be denoted by T, where $T = 1$ corresponds to following a high PEEP regime, and $T = 0$ corresponds to following a low PEEP regime. The outcome variable will be denoted by Y, where $Y = 1$ signifies that the patient died within 28 days, and $Y = 0$ signifies that the patient survived beyond 28 days.

Causal inference is important in this research as it helps to understand the cause-effect relationship between the treatment and the outcome. Causal effects refer to the difference in outcomes when comparing the result of giving treatment versus not giving treatment [11]. The notations $Y_i(0)$ and $Y_i(1)$ represent the potential outcomes for patient i under treatment $T = 0$ and $T = 1$, respectively. T is said to have a causal effect on patient i if $Y_i(0) \neq Y_i(1)$. There is no causal effect when $Y_i(0) = Y_i(1)$. The individual treatment effect (ITE) is defined as the difference between potential outcomes, representing the causal effect of treatment on the outcome for patient i [11, 12]:

$$ITE = Y_i(1) - Y_i(0)$$

Depending on the treatment given, only $Y_i(0)$ or $Y_i(1)$ can be observed, but not both [11]. This limitation is known as the fundamental problem of causal inference [8]. The other potential outcome remains unobserved and is considered missing data, making it challenging to exactly quantify the ITE. Instead, we can make use of the average treatment effect (ATE), which is defined as follows [8]:

$$ATE = E[Y(1)–Y(0)]$$

To estimate the effect of a treatment on a patient with certain patient characteristics x, the conditional ATE (CATE) can be used. The CATE function is defined as follows [8]:

$$\tau(x) = E[Y(1)–Y(0)|X = x]$$

2.4 Confounding

Due to the lack of randomization in observational studies, exchangeability cannot be assumed, as the treatment and control groups may not be comparable [11]. To address this limitation, we employ a strategy that treats the dataset as if the treatment assignment were randomized, conditioned on a measured set of confounders L. Confounders are variables that have an effect on both the treatment assignment and the outcome. By accounting for these confounders, we can identify the direct effect of the treatment on the outcome. We can consider the observational data as conditionally randomized (experimental) data if the following identifiability conditions hold [11, 13]:

- **Consistency**: The potential outcome under the treatment received is equal to the observed outcome. That is, $Y(T) = Y$.
- **Conditional exchangeability**: The treatment assignment is independent of the potential outcomes, given the measured confounders L. Formally, $Y(0), Y(1) \perp\!\!\!\perp T \mid L$.
- **Positivity**: For each combination of confounders L, there must be a non-zero possibility of receiving each treatment.

3 Methodology

To address both the main research question and the subquestion, this research adopts an experimental approach. The experiment involves applying the meta-learners to simulated data, the MIMIC-IV dataset, and an RCT dataset for external validation. In the subsequent sections, the meta-learners will be described in detail, along with the confounders identified in the MIMIC-IV dataset.

3.1 Meta-learners for CATE estimation

A meta-learner combines supervised learning or regression estimators (base learners), thereby enabling flexibility in the types of these base learners [8]. To verify the hypothesis regarding the X-learner, and to assess its performance in estimating the CATE, we compare the X-learner with the S- and T-learner. These two learners are commonly used in CATE-estimation problems [8], making them a straightforward choice in the comparison process.

S-learner

The S-learner uses a *single* machine-learning model M (which can be any model) to estimate the combined response function. This response function includes the treatment variable in the feature vector and is defined as follows:

$$\mu(x, t) = E[Y|X = x, T = t]$$

Using $\hat{\mu}(x, t)$, predictions can be made under different treatment assignments. The estimated CATE is then calculated as the difference between the predicted outcome under $T = 1$ and $T = 0$:

$$\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$$

T-learner

The T-learner uses *two* models M_0 and M_1 to estimate the response functions μ_0 (using the observations from the control group) and μ_1 (using the observations from the treatment group), respectively:

$$\mu_0(x) = E[Y|X = x, T = 0]$$
$$\mu_1(x) = E[Y|X = x, T = 1]$$

After training these models, the CATE estimation is calculated as follows:

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

X-learner

The X-learner can be described in three steps:

1. Similar to the T-learner, the first step is to estimate the response functions μ_0 and μ_1 using any two machine learning models M_0 and M_1:
$$\mu_0(x) = E[Y|X = x, T = 0]$$
$$\mu_1(x) = E[Y|X = x, T = 1]$$
2. Next, $\hat{\mu}_0$ and $\hat{\mu}_1$ are used to impute the treatment effect for the control and treatment groups as follows:
$$\hat{\tau}_0(x) = \hat{\mu}_1(x) - Y \quad \text{for } T = 0$$
$$\hat{\tau}_1(x) = Y - \hat{\mu}_0(x) \quad \text{for } T = 1$$
Two more models $M_{\tau 0}$ and $M_{\tau 1}$ (the second-stage models) are fit to estimate these imputed treatment effects:
$$\hat{\tau}_0(x) \sim M_{\tau 0}$$
$$\hat{\tau}_1(x) \sim M_{\tau 1}$$
3. Finally, the second-stage models are combined with a propensity score model $e(x)$ to estimate the CATE as follows:
$$\hat{\tau}(x) = \hat{e}(x)\hat{\tau}_0(x) + (1 - \hat{e}(x))\hat{\tau}_1(x)$$

The propensity score $e(x)$ indicates the probability of a unit receiving treatment $T = 1$, based on its covariate values [14]. When there are very few treated units, $\hat{M}_{\tau 0}(X)$ will be less accurate since the imputation is based on only a few samples. The X-learner handles this by assigning a low weight to the inaccurate model because $\hat{e}(x)$ will generally be low in this scenario. Conversely, $\hat{M}_{\tau 1}(X)$ will be more accurate since its imputation is based on a large sample set, thus a higher weight will generally be assigned to this model using $1-\hat{e}(x)$.

Base models

The performance of the S-, T-, and X-learners is highly dependent on the base model(s) used [12]. Therefore, different base models are used, including gradient boosted trees (specifically, LGBM), linear regression (LR), random forest (RF), and support vector regression (SVR). These models were chosen for their different levels of flexibility and complexity, providing a more thorough evaluation of the performance of the meta-learners.

Additionally, for the propensity score within the X-learner, different models are used. For simulated data, logistic regression (a commonly used model for the propensity score [12]) is used to focus on comparing different base models in estimating the CATE for unbalanced and confounded data. For the MIMIC-IV dataset, different models are evaluated for the propensity score to determine the most suitable approach for this specific dataset.

3.2 Set of confounders L

The used meta-learners can handle confounding, provided that the exchangeability assumption holds [8]. Therefore, all confounders present in the MIMIC-IV dataset must be correctly identified. We use feature selection together with literature reviews to derive this set of confounders. Using Scikit-learn [15], multiple methods are applied, including correlation analysis, univariate feature selection, recursive feature elimination, and tree-based feature selection (these results can be found in Appendix A). The combined results of this feature selection are shown in Tables 1 and 2.

Using literature reviews enables the identification of variables previously defined as confounders in similar studies. A meta-analysis comparing high versus low PEEP regimes showed that patients with a PF-ratio below a certain value benefit more from a high PEEP regime in terms of mortality [7]. This suggests that the PF-ratio influences the treatment assignment and outcome. The PF-ratio denotes the PaO2:FiO2 ratio [7]. Therefore, we consider the PF-ratio, PaO2, and FiO2 potential confounders. The same research hypothesized that patients with a higher body mass index (weight/height) benefit less from a high PEEP regime. Tables 1 and 2 show that weight is very predictive of both the treatment and outcome, but height is not. Therefore, only weight will be included in L.

A randomized trial by Meade et al. [16] comparing high versus low PEEP regimes states: "Protocols for reducing PEEP levels in the setting of hypotension (mean arterial pressure <60 mm Hg), high plateau airway pressures (<40 cm H_2O), or refractory barotrauma (see below) allowed us to further modify PEEP levels according to individual patient needs." This indicates that the PEEP regime is modified based on the plateau pressure. High plateau pressures indicate either worse disease severity or insufficient expansion of the lungs, and they independently contribute to a higher risk of mortality [17]. Therefore, we consider plateau pressure to be a confounder.

A Viewpoint article by Bugedo et al. [18] on driving pressure in ARDS patients mentioned that driving pressure is highly correlated with the survival outcome [18]. Besides, they suggest that the driving pressure may be a valuable measurement for setting the PEEP. Since driving pressure is hypothesized to affect both treatment and outcome, we consider it a confounder as well.

Additionally, age is very predictive of the outcome as shown in Table 2. Since age is a well-known indicator of mortality for critically ill patients, and therefore a big influence on the outcome, it is important to add it to set L. Similarly, it is shown that both HCO3 and respiratory rate are very predictive of the outcome variable. For the same reasoning, it is added to L. Besides that, PEEP is very predictive of the treatment assignment as shown in Table 1. However, since it is not predictive of the outcome, PEEP will not be included in L.

The other features will not be discussed as they are not in the validation set or do not have a significant association with either the treatment or the outcome according to our feature selection. Therefore, L contains the following features: age, weight, PF-ratio, PaO2, driving pressure, FIO2, HCO3,

plateau pressure, and respiratory rate.

Table 1: Combined results of the feature selection on the treatment variable.

Index	Feature
1	PEEP
2	Plateau pressure
3	Weight
4	FiO2
5	Age
6	PF-ratio
7	Respiratory rate
8	Minute volume
9	PaCO2
10	PaO2

Table 2: Combined results of the feature selection on the outcome variable.

Index	Feature
1	Age
2	Urea
3	Weight
4	HCO3
5	Respiratory rate
6	PaO2
7	Creatinine
8	Bilirubin
9	PaCO2
10	Lung Compliance

4 Experimental Setup and Results

The experiment conducted in this research consists of three parts: simulating data, analyzing the MIMIC-IV dataset, and performing external validation. The following sections describe these experiments and the results that were obtained.

4.1 Simulations

When assessing the performance of the meta-learners on real-world data, it is challenging to determine the accuracy of the model due to the absence of ground truth values. Using simulated MIMIC-like data allows us to compute the actual CATE, as both potential outcomes are simulated. This enables us to measure performance in terms of mean squared error (MSE) by comparing the estimated CATE to the simulated actual CATE.

Experiment

First, we generated MIMIC-like data using the simulation framework by Künzel et al. [8]. For this simulation, we needed to specify the dimension d (of the feature vector), the response functions $\mu_0(x)$ and $\mu_1(x)$, and the propensity score $e(x)$. We combined simulations 1 and 6 to create a simulation including unbalanced and confounded data. This was useful since the real-world MIMIC data is also unbalanced and confounded.

- **Feature vector:** We first simulated a d-dimensional feature vector X using $x \sim \text{Unif}([0,1]^{n \times d})$ with $d = 24$.
- **Potential outcomes:** We created potential outcomes using:

$$Y_i(0) = \mu_0(X_i) + \varepsilon_i(0)$$

$$Y_i(1) = \mu_1(X_i) + \varepsilon_i(1),$$

  where $\varepsilon_i(1), \varepsilon_i(0) \sim \mathcal{N}(0,1)$.
- **Response functions:** Simulation 6 shows that $e(x)$, $\mu_0(x)$ and $\mu_1(x)$ are based only on the confounding variable x_1 [8]. Since nine confounders were identified in the MIMIC-IV data, nine variables from the simulated feature vector X were selected randomly to serve as the confounders. An extra variable was added to μ_1 since in real life the outcome can depend on more than only the set of confounders. For each variable in the response functions, an arbitrary weight β_i was assigned as follows:

$$\begin{aligned}\mu_0(x) =& \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_4 + \beta_4 x_5 + \beta_5 x_9 \\ &+ \beta_6 x_{15} + \beta_7 x_{16} + \beta_8 x_{20} + \beta_9 x_{21} - 1\end{aligned}$$

$$\mu_1(x) = \mu_0(x) + \beta_{10} x_{19}$$

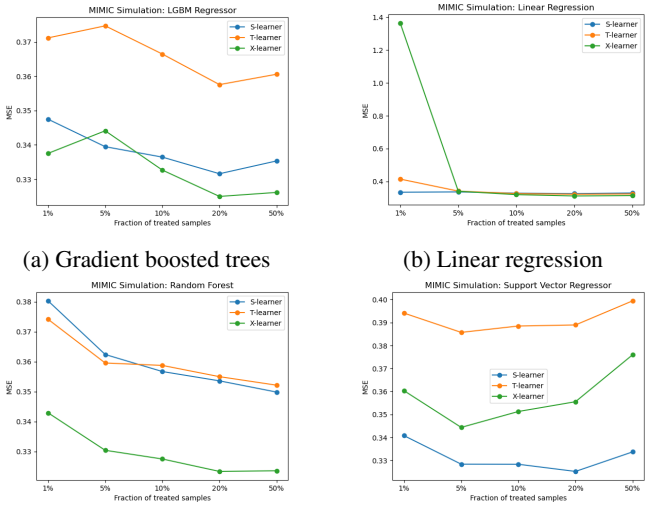
- **Propensity score:** The propensity score $e(x)$ was calculated using the logistic function applied to the linear combination of confounders L with randomly chosen coefficients. Additionally, it was important to adjust the propensity score to reflect the desired fraction of treated units. $e(x)$ and e_{adjusted} were defined as follows:

$$e(x) = \frac{1}{1 + e^{-L^T \beta}}$$

$$e_{\text{adjusted}} = \left(\frac{\text{treated fraction}}{\text{mean}(e)} \right) \cdot e(x)$$

- **Treatment assignment:** We simulated the treatment assignment according to $T_i \sim Bern(e(X_i))$, to obtain (X_i, T_i, Y_i), with $Y_i = Y(T_i)$. Finally, the actual CATE was calculated using $Y_i(1) - Y_i(0)$.

For the evaluation, we created simulations with propensity scores of 1%, 5%, 10%, 20%, and 50%. For each simulation, 4000 samples were generated and split into training and testing data (70% training, 30% testing). We trained each meta-learner on the training set and evaluated its performance against the test set using the actual CATE. This process was repeated 30 times (since the train and test set can differ for each run) to gain the average MSE for different propensity scores. For each of the underlying base models, we applied hyperparameter tuning to gain the best results in terms of MSE.



(a) Gradient boosted trees

(b) Linear regression

(c) Random forest

(d) Support vector regression

Figure 1: MSE of different simulations with varying propensity scores.

Results

The results from the experiments are shown in Figure 1. When using LGBM, the X-learner consistently has the lowest MSE, except at 5%, where the S-learner performs slightly better with a difference of 0.004 (which is negligible). For linear regression, the X-learner starts with a 'high' MSE for a treated sample fraction of 1%. From 5% onwards, the X-learner aligns with the stable low MSE of the S-learner and T-learner. The meta-learners perform similarly due to the linear relationship assumed by linear regression, resulting in comparable performance. Random forest consistently shows that the X-learner slightly outperforms the other learners. When using support vector regression, the S-learner seems to perform slightly better.

4.2 MIMIC-IV analysis

To evaluate the performance of meta-learners on MIMIC-IV data, we use Cumulative Gain Curves (CGC). These curves provide insight into how effectively models identify who benefit most from a treatment [12]. Initially, data is sorted by predicted CATE values in descending order. At each percentile of the sorted data, the cumulative gain is computed by measuring the cumulative treatment effect and the outcome across increasing proportions of the population. In our context, where outcome 1 represents undesirable outcomes (e.g., mortality) and outcome 0 represents favorable outcomes (e.g., survival), the cumulative gain at each percentile measures the overall improvement in identifying patients who are predicted to benefit from treatment compared to those who are not. The resulting curve is then compared to a random curve generated by a random model. The further the CGC is above this random curve, the better the model's performance.

The performance of the propensity score model within the X-learner will be evaluated in terms of accuracy and calibration. For the latter, we use the calibration curve and the Brier score [15]. Calibration curves can be used to compare the predicted probabilities of the propensity score model to the actual probabilities. A perfectly calibrated model will have points lying on the diagonal line on the plot. The Brier score measures the accuracy of the probabilistic predictions, with lower scores indicating higher accuracy.

Experiment

We started by pre-processing the MIMIC-IV dataset. The categorical feature 'sex' was converted to a numerical form ('F' to 0, 'M' to 1), as well as the 'peep_regime' feature ('low' to 0, 'high' to 1). Additionally, the 'mort_28' feature was converted from Boolean to numerical (False to 0, True to 1).

Since there were some missing values we had to apply imputation, we compared iterative imputation with k-means imputation (for k=2, k=6, k=12). For this, all samples in the dataset were selected without any missing values. In the resulting data frame, some values were randomly removed. Next, the imputation methods were applied to this dataset. The mean and standard deviation of the difference between the original data values and the imputed data values were calculated, these results can be found in Appendix B. From this, we concluded that for most features iterative imputation showed the best performance. Therefore, iterative imputation was used for the missing values in the original MIMIC-IV dataset.

After pre-processing the data, the MIMIC-IV data was split into a 70% training set and a 30% test set. Then, the data was normalized using the MinMaxScaler from Scikit-learn [15], scaling each feature to a value between 0 and 1. Next, we applied the S-, T-, and X-learner on this data (using only feature set L), to gain the predicted CATE for both the training and testing data. We computed the cumulative gain curve and area under the curve (AUC) for each meta-learner (using different base models). This process was repeated 100 times to gain the average cumulative gain curve and AUC. For each model, we applied hyperparameter tuning to improve the performance.

We applied more of an iterative approach for the X-learner by tuning the propensity score model and the second stage model. For the propensity score model, we analyzed various models, including logistic regression (LR), random forest classifier (RF), and decision tree classifier (DT). Again, we split the MIMIC-IV data into 70% training and 30% testing and normalized the data. Then, we trained the different models on the training set after which we predicted the treatment assignment probabilities for the test set. Next, we calculated the accuracy in terms of AUC-ROC, precision, recall, and F1-score using Scikit-learn [15]. Additionally, we calculated the calibration curve and the Brier score. This process was repeated 50 times to gain the average result. For each model, we applied hyperparameter tuning to improve the performance.

Assuming stage 1 handled confounding for the X-learner, we use a different set of features to fit the second-stage models. This is because some variables might significantly affect the outcome, but they do not affect the treatment variable, so they are not in L. The results of the previously mentioned feature selection on the outcome variable were used as features in the second-stage models.

Results

Table 3 shows that for the propensity score, logistic regression outperforms the other models. The calibration curve for using logistic regression on the MIMIC data, shown in Figure 2, suggests that the model is well-calibrated with a low Brier score. Therefore, the propensity model predicts the treatment probabilities quite accurately.

Table 3: Performance of different propensity score models in terms of accuracy.

Metric	LR	RF	DT
AUC-ROC	0.8272	0.7210	0.5278
Precision-score	0.4729	0.4374	0.1854
Recall-score	0.4296	0.1944	0.3265
F1-score	0.3894	0.2148	0.1908

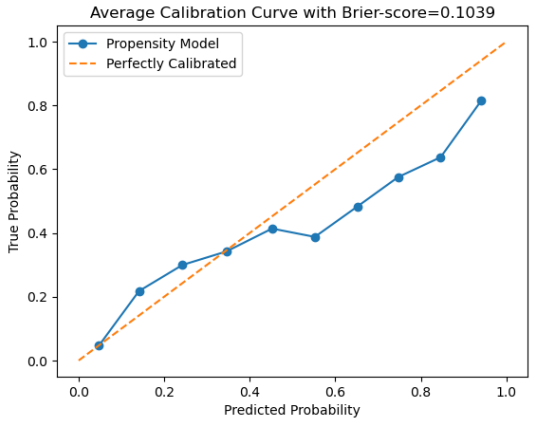


Figure 2: Averaged calibration curve for the propensity score model using logistic regression.

The results from applying the meta-learners to the MIMIC-IV data are shown below (for the cumulative gain curves see Appendix C). Tables 4 and 6 show that there is a large difference in the training and testing scores, which indicates that LGBM and RF are overfitting on the training data. Additionally, the curves of the testing set are very close to the random curve as indicated by the low mean AUC scores. Therefore, LGBM and RF may not be the best choices for CATE estimation on the MIMIC-IV data.

For LR, the train and test curves are more similar to each other. However, the AUC scores are very low for the training and testing curves and their standard deviations are quite high (see Table 5). This may be because linear regression assumes a linear relationship between the features and the outcome variable, which might be too simple for the MIMIC-IV data. SVR shows somewhat better performance, and the train and test curves do not differ significantly, suggesting that the model is not overfitting (see Table 7). The AUC is somewhat higher than for LR, suggesting better generalization and performance, but this difference is not significant. Overall, these results indicate the ability of the meta-learners to find a subgroup of patients that benefit from high PEEP treatment remains questionable. Notably, the AUC scores show that the X-learner is slightly outperforming the S-, and T-learner on the test set. This might be an indication of the X-learner performing better under the unbalanced MIMIC-IV dataset. However, due to the low mean AUC scores and the high standard deviations, this hypothesis cannot be verified.

Table 4: AUC scores using LGBM as base model(s).

Learner	Mean	SD
S-learner (test)	0.90	1.49
T-learner (test)	0.29	1.19
X-learner (test)	0.69	1.38
S-learner (train)	12.18	1.65
T-learner (train)	21.00	1.13
X-learner (train)	15.10	1.30

Table 5: AUC scores using LR as base model(s).

Learner	Mean	SD
S-learner (test)	0.36	1.23
T-learner (test)	0.85	1.31
X-learner (test)	0.92	1.32
S-learner (train)	0.20	0.79
T-learner (train)	2.78	0.98
X-learner (train)	2.33	0.69

Table 6: AUC scores using RF as base model(s).

Learner	Mean	SD
S-learner (test)	0.44	1.34
T-learner (test)	0.15	1.26
X-learner (test)	1.06	1.75
S-learner (train)	25.22	1.18
T-learner (train)	27.55	0.40
X-learner (train)	20.93	1.36

4.3 External validation

Experiment

For the external validation, the meta-learners were trained on the entire MIMIC-IV dataset, using SVR as the base model. The same preprocessing steps as for the MIMIC-IV data were applied to the RCT dataset. Subsequently, the trained models were applied to the RCT data to obtain the predicted CATE values.

Results

The cumulative gain curves for the S-, T-, and X-learners on the RCT dataset are shown in Figure 3. The curves are quite close to the baseline, suggesting that the CATE predictions do not significantly differ from those of a random model. This

Table 7: AUC scores using SVR as base model(s).

Learner	Mean	SD
S-learner (test)	2.90	1.66
T-learner (test)	2.81	1.79
X-learner (test)	2.98	1.46
S-learner (train)	4.34	1.02
T-learner (train)	4.47	0.92
X-learner (train)	4.17	0.78

indicates that the models are not effectively distinguishing between high PEEP and low PEEP groups. In contrast to the MIMIC-IV data, where the cumulative gain curves ascend, the curves for the RCT data are descending. From the cumulative gain curves, it can be derived that for the RCT data, we cannot find a subgroup of patients that benefit from high versus low PEEP using the S-, T-, and X-learner trained on the observational MIMIC-IV data. Additionally, none of the meta-learners are outperforming each other on the RCT data.

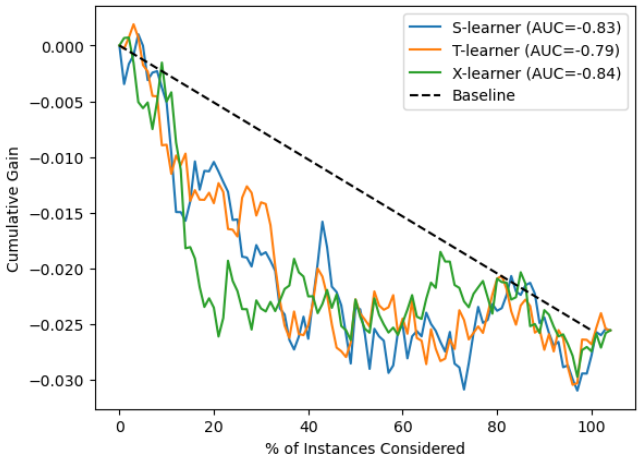


Figure 3: Cumulative gain curve for the RCT dataset.

5 Discussion

From the simulation results, it was found that the X-learner performed well in terms of MSE when using LGBM, LR, and SVR, across various unbalanced conditions. These findings verify the results of previous studies showing the effectiveness of the X-learner under unbalanced treatment and control groups. Additionally, the overall high performance of the meta-learners indicates their ability to accurately estimate the CATE, even under confounding conditions.

Despite the use of different iterations of the experiment to gain an averaged result, the conclusions were drawn from a single simulation setup. In this setup, the response functions were linear. In future research, it would be valuable to investigate the performance of non-linear response functions as well. Besides that, the simulated data is only an approximation of the MIMIC-IV data. It cannot fully replicate the complexity of the real-world MIMIC-IV dataset. Therefore, we should not fully depend on the performance of the CATE estimators on the simulation data.

For the MIMIC-IV analysis, the performance of the meta-learners on real-world observational data was examined. When using LGBM and RF, the meta-learners were overfitting on the training data. A possible explanation for this is that LGBM and RF are quite complex models, which may fit the training data too closely. This can result in the models capturing noise and not generalizing well to new, unseen data. SVR performed best in terms of AUC and showed the lowest difference between the training and testing curves.

For all base models, the meta-learners did not show significant performance on the MIMIC-IV dataset. The curves corresponding to the test set were close to the curve resembling the random model. Therefore, the meta-learners were unable to identify a subgroup of patients that may benefit from high PEEP. This might be because the meta-learners were unable to model the complex relationships underlying the MIMIC-IV data, though other external factors could have influenced the CATE estimates.

There is a possibility that some confounding variables were overlooked, leading to hidden confounding factors. These unobserved confounders might have influenced the CATE estimations, potentially introducing bias into the results. Additionally, from the feature selection, it was shown that the PEEP variable is highly influential in the prediction of the treatment assignment. Including this variable in the propensity score model of the X-learner significantly increased its performance in terms of accuracy and calibration. However, including PEEP in the propensity score model would partially reveal the treatment assignment. Therefore, the propensity score model did not include this variable, resulting in the model performing somewhat worse but without any bias. While the X-learner showed a slightly better performance in terms of AUC compared to the S- and T-learners on the MIMIC-IV dataset, the high standard deviations among the AUC values indicate significant uncertainty in the results. Therefore, it cannot be confidently concluded that the X-learner performs well under unbalanced data.

For the external validation, the results show that the models trained on the MIMIC-IV data perform worse on the RCT dataset. This may be because the meta-learners fail to correctly predict the CATE estimates and identify a subgroup of patients that benefit from high PEEP. There may also be some issues outside the models causing these results.

From the feature selection on the MIMIC-IV, it was shown that there are several features highly influential on the outcome variable but not present in the RCT dataset. These features were excluded in training the models for the RCT dataset, which may have caused the models to perform worse. Additionally, the distribution of the data in the RCT dataset and the MIMIC-IV dataset might differ significantly. This may have caused the models to perform almost randomly on the RCT-data, due to overfitting on the MIMIC-IV data. Besides that, the treatment assignment in the RCT dataset was balanced, with 49% of the patients treated. This means the meta-learners were trained on unbalanced data and then applied to balanced data. As a result, the CATE estimates might be biased towards underestimating the treatment effect for the

treated group in the MIMIC-IV data, potentially leading to inaccurate results for the treatment effect in the RCT dataset.

6 Responsible Research

Several ethical aspects must be considered when conducting research in the medical field to ensure responsible research practices. A significant ethical concern is the potential for bias and unfairness in treatment assignments. The CATE estimators are trained on the specific MIMIC-IV data, which may include some bias. Depending entirely on these treatment assignments and corresponding outcomes could lead to unfairness for patients outside the MIMIC-IV dataset, as they may not receive appropriate treatment based on these estimators. Consequently, these estimators (trained on limited data) might overfit to this dataset, which raises ethical concerns regarding fairness to other patients.

Additionally, the use of machine learning models in healthcare needs some additional responsibility to be used properly. Misuse or over-reliance on the recommendations without any clinical oversight may lead to adverse patient outcomes. Therefore, these types of models should only be used to give a suggestion and then be confirmed by healthcare professionals before being used in practice.

To facilitate the reproducibility of the methods used, all code developed for this research is made publicly available in a GitHub repository.[1] This includes the implementation of the S-, T-, and X-learners, the simulations, the MIMIC-IV analysis, and the external validation steps. The code includes detailed documentation to guide other researchers through the experimental steps. Note that access to the MIMIC-IV dataset and the RCT data is required to reproduce the experiments involving these datasets.

7 Conclusions and Future Research

This paper aimed to answer the question whether the S-, T-, and/or X-learner can be used to predict which ICU patients suffering from ARDS benefit from high PEEP compared to low PEEP in mechanical ventilation based on patient characteristics. Additionally, we aimed to verify the hypothesis regarding the X-learner performing particularly well under unbalanced data (in terms of treatment assignment). An experimental approach was conducted by applying the S-, T-, and X-learner (using different base models) to estimate the CATE for simulated data, real-world MIMIC-IV data, and data from a randomized control trial for external validation. Due to the fundamental problem of causal inference, it was necessary to identify the confounders present in the real-world dataset by doing literature reviews and applying feature selection methods.

We generated simulation data, with confounding present and different treatment assignment distributions ranging from 1% to 50%. The meta-learners had a high performance on this simulation data in terms of MSE, using different base models including gradient boosted trees (specifically, LGBM), linear regression (LR), random forest (RF), and support vector regression (SVR). For the application of the meta-learners on the MIMIC-IV data, the same base models were used. Based on these results, it seems that LGBM and RF are overfitting on the training data. Overall, the S-, T-, and X-learners were not able to identify a subgroup of patients that benefit from high PEEP compared to low PEEP, as indicated by the low AUC scores in the cumulative gain curves and the high variability of these results. After training the meta-learners on the entire MIMIC-IV dataset using SVR, they were applied to an external RCT dataset. The cumulative gain curves showed that all learners performed worse than the random baseline model. This suggests that for the RCT data, the S-, T-, and X-learners cannot identify a subgroup of patients that benefit from high versus low PEEP. Additionally, the experiments showed that the X-learner slightly outperformed the S-, and T-learner under unbalanced data. However, due to the variability of the results and because of the small difference in performance, the hypothesis regarding the X-learner cannot be verified.

To further investigate the hypothesis that some patients benefit more from high versus low PEEP using the S-, T-, and X-learner, additional research is needed. Similarly, more research is needed in the future to verify the hypothesis about the X-learner performing particularly well under unbalanced treatment assignment. It might be useful to examine a broader range of base models (e.g., neural networks). Additionally, it might be interesting to experiment with combining different base models for the T-, and X-learner. For simulating data, it is recommended to use varying, more complex response functions. Besides that, future research could look into methods to mitigate the impact of potentially hidden confounding variables to improve the reliability of the CATE estimates. For the external validation, it might be useful to verify whether the external data is distributed similarly to the training data.

[1] https://github.com/JuulSchnitzler/Estimating-CATE-using-the-meta-learners

References

[1] M. J. Tobin, "Advances in mechanical ventilation," *New England Journal of Medicine*, vol. 344, no. 26, pp. 1986–1996, Jun. 2001.

[2] A. J. Walkey, L. Del Sorbo, C. L. Hodgson, N. K. J. Adhikari, H. Wunsch, M. O. Meade, E. Uleryk, D. Hess, D. S. Talmor, B. T. Thompson, R. G. Brower, and E. Fan, "Higher peep versus lower peep strategies for patients with acute respiratory distress syndrome: A systematic review and meta-analysis," *Annals of the American Thoracic Society*, vol. 14, pp. S297–S303, 2017.

[3] Y. Oba, D. M. Thameem, and T. Zaza, "High levels of peep may improve survival in acute respiratory distress syndrome: A meta-analysis," *Respiratory Medicine*, vol. 103, no. 4, pp. 585–591, 2009.

[4] A. B. Cavalcanti, É. A. Suzumura, L. N. Laranjeira, D. de Moraes Paisani, L. P. Damiani, H. P. Guimarães, E. R. Romano, M. d. M. Regenga, L. N. Takahashi Taniguchi, C. Teixeira, R. P. d. Oliveira, F. R. Machado, F. A. Diaz-Quijano, M. S. de Alencar Filho, I. S. Maia, E. B. Caser, W. de Oliveira Filho, M. d. C. Borges, P. d. A. Martins, M. Matsui, G. A. Ospina-Tascón, T. S. Giancursi, N. D. Giraldo-Ramirez, S. R. R. Vieira, M. d. G. P. d. L. Assef, M. S. Hasan, W. Szczeklik, F. Rios, M. B. P. Amato, O. Berwanger, and C. R. Ribeiro de Carvalho, "Effect of lung recruitment and titrated positive end-expiratory pressure (peep) vs low peep on mortality in patients with acute respiratory distress syndrome: A randomized clinical trial," *JAMA*, vol. 318, no. 14, pp. 1335–1345, Oct. 2017.

[5] A. Serpa Neto and M. J. Schultz, "Optimizing the settings on the ventilator: High peep for all?" *JAMA*, vol. 317, no. 14, pp. 1413–1414, 2017.

[6] C. S. Calfee, K. Delucchi, P. E. Parsons, B. T. Thompson, L. B. Ware, and M. A. Matthay, "Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials," *The Lancet Respiratory Medicine*, vol. 2, no. 8, pp. 611–620, Aug. 2014.

[7] M. Briel, M. Meade, A. Mercat, R. G. Brower, D. Talmor, S. D. Walter, A. S. Slutsky, E. Pullenayegum, Q. Zhou, D. Cook, L. Brochard, J.-C. M. Richard, F. Lamontagne, N. Bhatnagar, and T. E. Stewart, "Higher vs lower positive end-expiratory pressure in patients with acute lung injury and acute respiratory distress syndrome: Systematic review and meta-analysis," *JAMA*, vol. 303, no. 9, pp. 865–873, 2010. [Online]. Available: https://doi.org/10.1001/jama.2010.218

[8] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 116, no. 10, pp. 4156–4165, Feb. 2019.

[9] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, and R. G. Mark, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10.

[10] R. G. Brower, P. N. Lanken, N. MacIntyre, M. A. Matthay, A. Morris, M. Ancukiewicz, D. Schoenfeld, and B. T. Thompson, "Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome," *New England Journal of Medicine*, vol. 351, no. 4, pp. 327–336, Jul. 2004.

[11] M. A. Hernán and J. M. Robins, *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2024.

[12] M. F. Alves. (2022) Python causality handbook. [Online]. Available: https://matheusfacure.github.io/python-causality-handbook/landing-page.html

[13] AI/ML Services, Australia. Assumptions. [Online]. Available: https://www.causalwizard.app/inference/article/assumptions

[14] C. U. M. S. of Public Health, "Propensity score analysis — columbia public health," Oct. 2022. [Online]. Available: https://www.publichealth.columbia.edu/research/population-health-methods/propensity-score-analysis

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] M. O. Meade, D. J. Cook, G. H. Guyatt, A. S. Slutsky, Y. M. Arabi, D. J. Cooper, A. R. Davies, L. E. Hand, Q. Zhou, L. Thabane *et al.*, "Ventilation strategy using low tidal volumes, recruitment maneuvers, and high positive end-expiratory pressure for acute lung injury and acute respiratory distress syndrome," *JAMA*, vol. 299, no. 6, p. 637, Feb. 2008.

[17] S. V. Chhangani, "Independent lung ventilation and bronchopleural fistula," in *Elsevier eBooks*, 2008, pp. 337–354.

[18] G. Bugedo, J. Retamal, and A. Bruhn, "Driving pressure: a marker of severity, a safety limit, or a goal for mechanical ventilation?" *Critical Care*, vol. 21, no. 1, Aug. 2017. [Online]. Available: https://doi.org/10.1186/s13054-017-1779-x

A Feature Selection

A.1 Treatment variable

Table 8: Top 10 features in terms of correlation with the treatment variable.

Feature	Correlation
PEEP	0.4991
Plateau pressure	0.3491
Weight	0.2794
FiO2	0.2349
Age	0.1801
PF-ratio	0.1755
Respiratory rate	0.1708
Minute volume	0.1703
PaCO2	0.1492
PaO2	0.1447

Table 9: Top 10 features in univariate feature selection on the treatment variable.

Feature	Scores
PEEP	1175.41
Plateau pressure	491.95
Weight	300.07
FiO2	207.03
Age	118.79
PF-ratio	112.56
Respiratory rate	106.45
Minute volume	105.83
PaCO2	80.63
PaO2	75.81

Table 10: Selected features from Recursive Feature Elimination (RFE) on the treatment variable.

Selected features (optimal number = 8)
Age
Weight
Lung compliance
Map
Heart rate
PEEP
Respiratory rate
Diastolic blood pressure

Table 11: Top 10 features using tree-based feature selection on the treatment variable.

Feature	Importance
PEEP	0.1599
Weight	0.0658
Plateau pressure	0.0610
FiO2	0.0521
Age	0.0448
PF-ratio	0.0401
pH	0.0372
Minute volume	0.0371
Respiratory rate	0.0369
PaO2	0.0360

Table 12: Top 10 features in terms of correlation with the outcome variable.

Feature	Correlation
Age	0.1526
Urea	0.1284
Weight	0.0957
HCO3	0.0922
Respiratory rate	0.0805
PaO2	0.0647
Creatinine	0.0646
Bilirubin	0.0606
PaCO2	0.0589
Lung compliance	0.0587

Table 13: Top 10 features using univariate feature selection on the outcome variable.

Feature	Scores
Age	84.44
Urea	59.45
Weight	32.78
HCO3	30.37
Respiratory rate	23.12
PO2	14.91
Creatinine	14.85
Bilirubin	13.05
PaCO2	12.32
Lung compliance	12.27

Table 14: Selected features from Recursive Feature Elimination (RFE) on outcome variable.

Selected features (optimal number = 10)
Age
Weight
PF-ratio
PaO2
Driving pressure
Lung compliance
Bilirubin
Urea
FiO2
Minute volume

Table 15: Top 10 features using tree-based feature selection on the outcome variable.

Feature	Importance
Age	0.0568
Urea	0.0506
Weight	0.0458
Bilirubin	0.0457
pH	0.0442
Heart rate	0.0437
HCO3	0.0431
Minute volume	0.0431
Creatinine	0.0430
Platelets	0.0425

B Imputation Methods

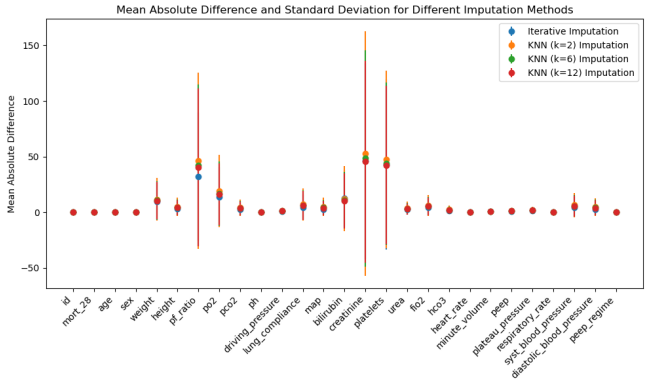


Figure 4: Mean absolute difference and standard deviation for different imputation methods.

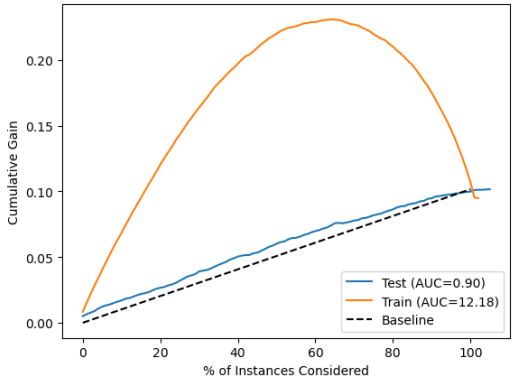
Table 16: Best performance in mean absolute difference.

Method	Count of Features
Iterative Imputation	24
KNN Imputation, k=2	0
KNN Imputation, k=6	0
KNN Imputation, k=12	3

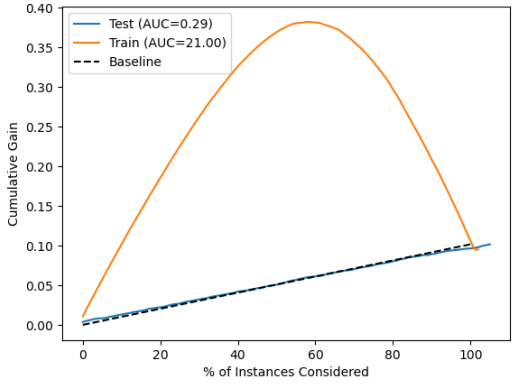
Table 17: Best performance in standard deviation

Method	Count of Features
Iterative Imputation	25
KNN Imputation, k=2	0
KNN Imputation, k=6	0
KNN Imputation, k=12	2

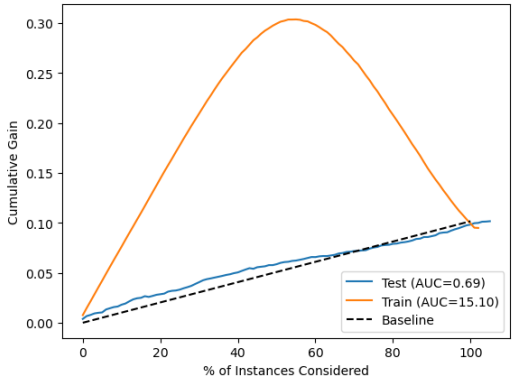
C MIMIC-IV Results



(a) S-learner

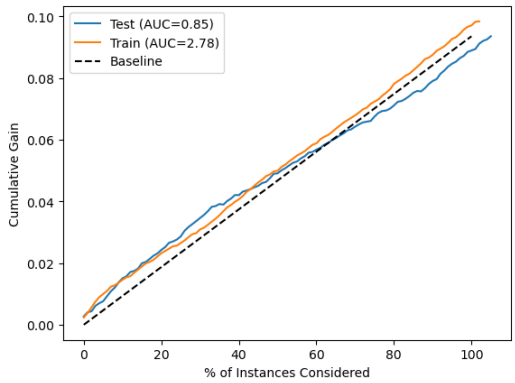


(b) T-learner

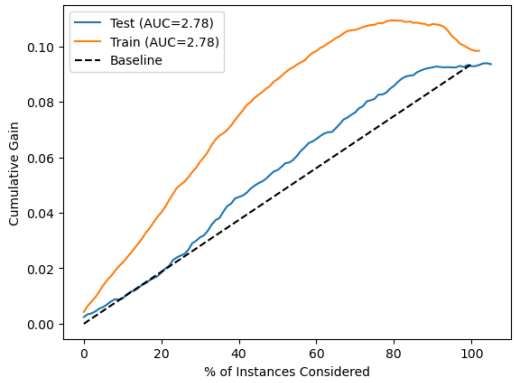


(c) X-learner

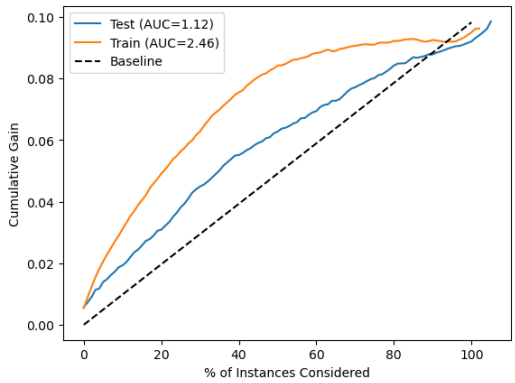
Figure 5: Average cumulative gain curves on training and testing set using LGBM.



(a) S-learner

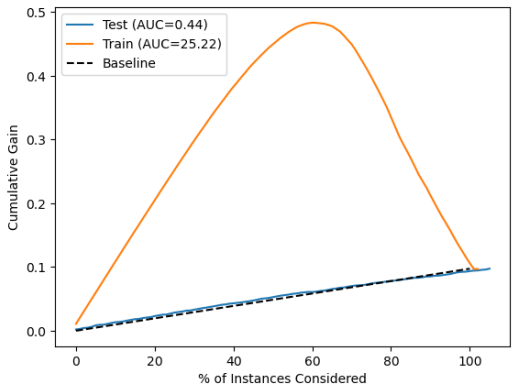


(b) T-learner

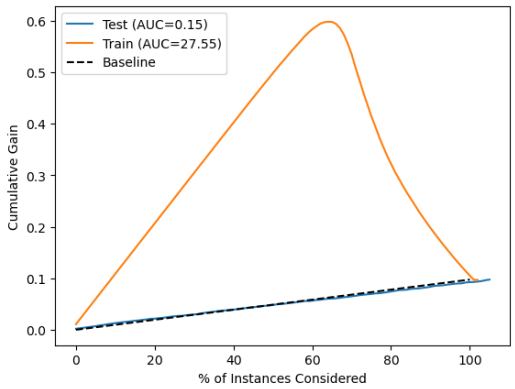


(c) X-learner

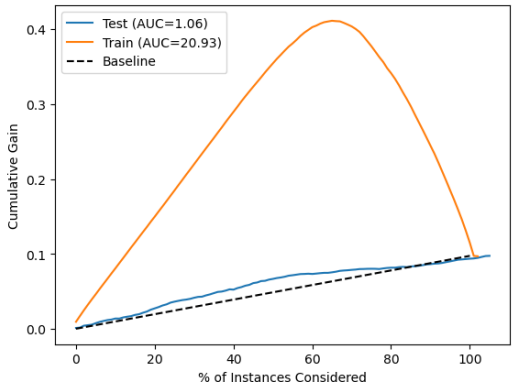
Figure 6: Average cumulative gain curves on training and testing set using linear regression.



(a) S-learner

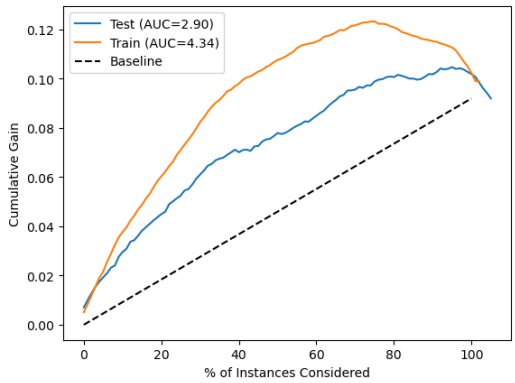


(b) T-learner

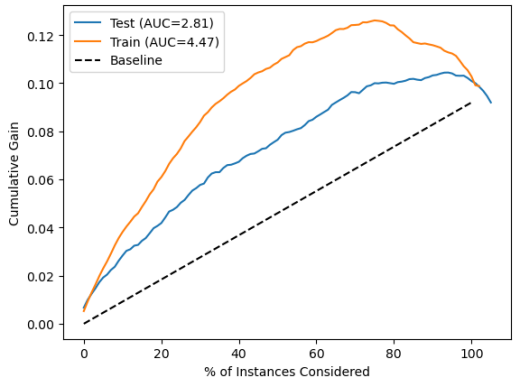


(c) X-learner

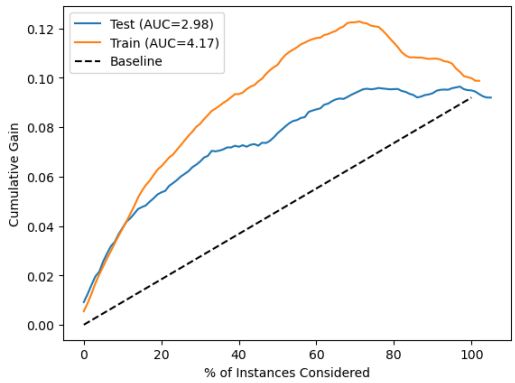
Figure 7: Average cumulative gain curves on training and testing set using random forest.



(a) S-learner



(b) T-learner



(c) X-learner

Figure 8: Average cumulative gain curves on training and testing set using support vector regression.