Machine Learning for Personalized Respiratory Care: A DR-learner Approach to Positive End-Expiratory Pressure Effect Estimation

Robert Melika[1]

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

[1]EEMCS, Delft University of Technology, The Netherlands
[2]Department of Intensive Care, Erasmus Medical Center, Rotterdam, 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: Robert Melika[1]
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe[1], Rickard Karlsson[1], Jim Smit[1,2], Jasmijn Baaijens[1]

Abstract

Mechanical ventilation with positive end-expiratory pressure (PEEP) is a critical intervention for patients in intensive care units (ICUs) with acute respiratory failure. Identifying the optimal PEEP level is challenging due to conflicting evidence from studies comparing low and high PEEP regimes. This research explores machine learning methods for estimating individualized treatment effects (ITE) in ICU patients on different PEEP levels using the observational MIMIC-IV dataset. Various conditional average treatment effect (CATE) estimators, including S-, T-, and DR-learners, are applied to control for confounders and identify PEEP effects on patient subgroups. This research aims to compare the performance of the aforementioned CATE estimators, with a focus on the doubly-robust (DR) learner, and determine which one is best suited for causal inference in this context. The DR-learner offers increased resilience to model errors since it integrates two models. Simulations using mean squared error (MSE) show the DR-learner performs well with confounded data and differing linear response functions between control and treatment groups. However, when looking at the performance on the MIMIC-IV dataset, the predictions are unstable, failing to reliably identify the optimal PEEP for increasing patient survival. This trend is also observed in a randomized controlled trial (RCT) dataset, with the area under the Qini curve (AUQC) close to zero, indicating difficulties in identifying the effects of PEEP settings. Despite promising simulation results, real-world application shows limitations in these machine learning methods for optimal PEEP identification.

1 Introduction

Mechanical ventilation in ICUs commonly employs positive end-expiratory pressure (PEEP) to support patients with acute respiratory failure. While PEEP helps prevent alveolar collapse and improves oxygenation, selecting the right level is complex. The literature presents conflicting evidence regarding the benefits of low and high PEEP regimes [1], which complicates treatment decisions.

To address this uncertainty, we would like to estimate the individualized treatment effect (ITE). The ITE, also known as the conditional average treatment effect (CATE), measures the effect of a treatment on an individual or a specific subgroup rather than on the overall population. Machine learning-based methods for ITE/CATE estimation can provide a valuable approach to personalizing medical treatments. These methods allow researchers to analyze large datasets and identify patient characteristics that influence the optimal PEEP regime. Despite this potential, there is a gap in understanding how best to apply these methods for causal inference in observational data. One major challenge is dealing with confounding variables, which affect both the treatment and the outcome and can significantly impact the results.

Our research aims to fill this gap by applying various CATE estimators to the MIMIC-IV dataset [2], an observational dataset that contains inherent confounding variables, to predict survival outcomes for ICU patients under different PEEP regimes. This paper will attempt to answer the following main research question: *How can the DR-learner, a machine learning-based method, be used to predict survival outcomes in ICU patients under different PEEP regimes based on individual characteristics, and how does this method compare to other CATE estimators when evaluated on an RCT dataset?* The objective is to compare the performance of different CATE estimators, more specifically the meta-learners S-, T-[3], and DR-[4], with a focus on the doubly-robust (DR) learner, to assess their robustness in handling confounding and providing accurate CATE estimates. The DR-learner should offer reduced bias due to its combination of two approaches. This research will attempt to contribute to personalized treatment strategies in ICUs by identifying patient subgroups that benefit most from low or high PEEP.

Firstly, the paper will focus on evaluating and comparing the meta-learners on simulated data using Gradient Boosting. Afterwards, the focus will switch to the MIMIC-IV dataset, where first imputation of missing data will be handled, and where the application of the different meta-learners will be executed with multiple underlying models: Linear Regression, Gradient Boosting, Support Vector Machines (SVMs), Logistic Regression, and eXtreme Gradient Boosting (XGBoost). Finally, the trained meta-learners will be evaluated on the MIMIC-IV dataset and on a randomized control trial (RCT) dataset.

2 Background and Terminology

This section outlines our objective of estimating the CATE using machine learning. We will introduce the dataset utilized in our research and discuss the broader issues of causal inference and confounding variables.

2.1 MIMIC-IV Dataset

We utilize the MIMIC-IV dataset, a comprehensive collection of ICU patient data, which contains 3,941 samples of patients with hypoxemic respiratory failure. As an observational dataset, MIMIC-IV contains inherent confounding variables, which must be carefully addressed to ensure accurate estimation and interpretation. Otherwise, the true effect of the treatment could be distorted. For our research, the dataset has been pre-processed to contain 26 covariates.

- Feature set X - 24 characteristics of a patient, such as age, heart rate, etc.
- Treatment variable - *peep_regime*, which is either *high* - **treatment** or *low* - **control**
- Outcome variable - *mort_28*, which is either *True* or *False*, is the mortality after 28 days

However, it is important to note that in this paper we look at the inverted outcome, *survival*.

Early analysis of the MIMIC-IV dataset tells us that approximately 12% of the patients were treated (received high PEEP).

Our goal is to estimate the CATE, more specifically to determine whether high PEEP is beneficial based on the characteristics of a patient using the MIMIC-IV dataset. The CATE, denoted as τ_i, represents the expected treatment effect for an individual with specific characteristics. While τ_i itself is impossible to observe directly due to the observational nature of the dataset and the presence of confounding variables, it can be estimated if certain assumptions hold. These assumptions are critical for making valid inferences from the data.

2.2 Causal Inference and CATE

Causal inference involves methods and techniques used to estimate the causal effect of a treatment. This process often requires taking confounding variables into account.

To formally define causal inference, we need to consider the potential outcomes framework. Let $Y_i(1)$ and $Y_i(0)$ represent the potential outcomes for an individual i under treatment and control conditions, respectively. The individual treatment effect (ITE) for person i is:

$$\tau_i = Y_i(1) - Y_i(0) \tag{1}$$

However, since we can only observe one of these outcomes for each individual (either $Y_i(1)$ or $Y_i(0)$, but not both), we rely on statistical methods to estimate the average treatment effect (ATE) or the conditional average treatment effect (CATE).

The CATE is a key concept in causal inference, especially for personalized medicine. It represents the average treatment effect for a subset of datapoints characterized by a specific set of features $X = x$. Mathematically, CATE is defined as:

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] \tag{2}$$

The goal of estimating CATE is to determine how the treatment effect varies across different subgroups defined by their covariates X. In our case, determining whether low or high PEEP is more beneficial in terms of increased chances of survival. This allows for more personalized and effective treatment strategies. A positive CATE indicates that high PEEP improves the patient's chances of survival, while a negative CATE suggests that low PEEP is more beneficial.

To accurately estimate the CATE, certain assumptions must be satisfied:

1. **Consistency:** The potential outcomes under treatment and control correspond to the observed outcomes when the individual receives the treatment or control respectively. Formally, if A_i is the treatment indicator, then $Y_i = Y_i(A_i)$.
2. **Positivity (Overlap):** Every individual has a positive probability of receiving both the treatment and the control, given their covariates. Formally defined, $0 < \mathbb{P}(A_i = 1 \mid X_i = x) < 1$ for all x.
3. **Conditional Exchangeability (No Unmeasured Confounding):** Given the covariates X_i, the treatment assignment is independent of the potential outcomes. Formally, $Y_i(1), Y_i(0) \perp A_i \mid X_i$.

These assumptions are essential to ensure that the estimates of CATE are unbiased and reliable, enabling us to make valid causal inferences from the observational data.

2.3 Confounding variables in the MIMIC-IV dataset

Confounding variables influence both the treatment assignment A and the potential outcomes $Y(1)$ and $Y(0)$. Since MIMIC-IV is an observational dataset, we can assume it is subjected to confounding. To identify the confounding variables, we have combined two approaches: data-driven analysis and literature review.

Data-Driven Approach

We first predict the outcome and treatment assignment separately to identify the most influential features. The propensity score and outcome predictions exclude the *peep* feature, as it is part of the treatment itself and does not provide conclusive information. For additional Figures, see Appendix A. We used the Gradient Boosting classifier from the *scikit-learn* library due to its ability to indicate feature importance.

Estimated outcome:

$$\eta(x) = \mathbb{E}[Y \mid X = x] \tag{3}$$

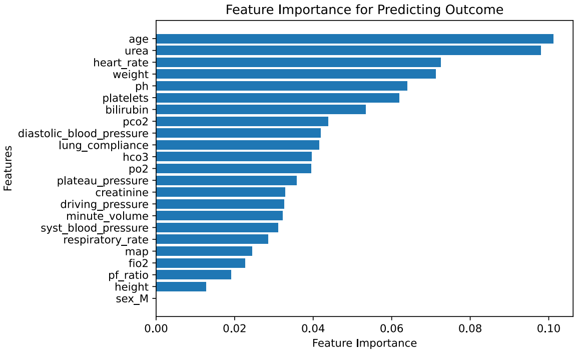


Figure 1: Feature importance when estimating outcome

Estimated treatment assignment:

$$\xi(x) = \mathbb{E}[A \mid X = x] \tag{4}$$

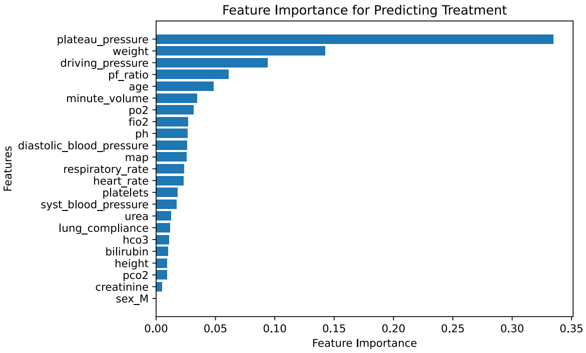


Figure 2: Feature importance when estimating treatment assignment

The features that overlap in the top 10 are *po2, age, ph, weight*. These could be potential confounders. We decided to also include outcome predictors (prognostic factors) in our set of confounding variables. Training on the outcome predictors additionally to the confounders does not diminish the performance of the estimator. However, training on the treatment predictors would, especially the propensity score [5]. The most significant outcome predictors are *age, urea, heart_rate, weight, ph, platelets*.

Literature review

To complement our data-driven approach, comprehensive literature analysis identified the following confounders:

- **PF ratio, PaO_2, FiO_2:** The PF ratio is a measure of the severity of respiratory failure and is used to assess the degree of hypoxemia in patients with acute respiratory distress syndrome (ARDS). A meta-analysis comparing high versus low PEEP regimes found that patients with a PF ratio (PaO_2/FiO_2) below 200 benefit more from a high PEEP regime in terms of mortality outcomes [6]. Therefore, *PF ratio, PaO_2, and FiO_2* are considered potential confounders.
- **Weight:** Body mass index (BMI) is a well-known factor influencing treatment outcomes in critical care settings. Research indicates that patients with a higher BMI may benefit less from high PEEP regimes due to altered respiratory mechanics and increased chest wall stiffness [6]. Our feature selection process also highlighted weight as a significant predictor for both treatment and outcomes, reinforcing its inclusion as a confounder.
- **Plateau Pressure:** Plateau pressure measures the pressure in the lungs during mechanical ventilation. When plateau pressure is high, it indicates more severe lung disease or poor lung expansion, both of which increase the risk of death [7]. Research shows that adjusting PEEP based on plateau pressure can improve outcomes [8]. A threshold of 29 cm H_2O is identified, above which the risk of death rises significantly [7].
- **Driving Pressure:** Driving pressure, defined as the difference between plateau pressure and PEEP, is highly correlated with survival outcomes and is used to set PEEP levels, making it a confounder. A cut-off value of 19 cm H_2O has been identified, above which the risk of death increases significantly. Therefore, driving pressure is a significant predictor of outcome [7].

Established Set of Selected Variables

Combining both data-driven and literature-based approaches, our final set of confounders and outcome predictors is composed of*Age, Weight, PF ratio, PO_2, Urea, pH, FiO_2, Plateau pressure and Driving pressure*.

3 Methodology

This section describes the experimental approach, the evaluation metrics and the use of meta-learners to estimate the CATE.

3.1 Experimental Approach

1. First, we will generate six simulated datasets following the methodology of Künzel et al. [3] with two additional simulations to compare the performance of selected meta-learners. By using simulated data, we can validate our methods against known ground truths, ensuring that our approach is sound before applying it to real-world data.
2. Next, we will pre-process the MIMIC-IV dataset, including scaling, imputing missing data and selecting variables. Controlling for confounders and outcome predictors will help us isolate the true effect of high PEEP on patient outcomes, aiding in result robustness.
3. We will then train the same selected meta-learners on the pre-processed MIMIC-IV dataset. Meta-learners are advanced machine learning models specifically designed to estimate the CATE. Training these models on real-world data should allow us to capture the complex relationships between patient characteristics, treatment, and outcomes.
4. Following training, we will evaluate the performance of the meta-learners using the Qini curve [9] and the areas under it (AUQC). These evaluation metrics will provide insights into how well each model estimates the CATE.
5. Finally, we will train the meta-learners again on the MIMIC-IV dataset, this time only with features that are available in the RCT dataset. Then we will again validate the CATE estimated on an RCT dataset using the Qini curve and the AUQC.

3.2 Qini Curve

To evaluate the performance of our trained meta-learners and CATE estimates, we will use the Qini curve and the AUQC. The Qini curve is a tool used in treatment effect estimation to assess treatment decisions, since it measures the incremental gain achieved by applying a treatment based on the estimated CATE scores.

First, the individuals are ranked by their estimated treatment effects, from highest to lowest. Then, the cumulative number of positive outcomes is plotted against the cumulative number of treated individuals, and a diagonal baseline represents the expected gain if treatment were assigned randomly. Finally, the area between the Qini curve and the baseline quantifies the model's ability to predict the treatment effect. A higher AUQC indicates better performance.

Interpretation

Positive slope indicates the model correctly identifies individuals who benefit most from the treatment. Comparing Qini curves across models shows which model provides the most accurate CATE estimates. The model with the highest Qini curve and AUQC would be considered the best performer.

3.3 Meta-learners

Meta-learners are machine learning techniques designed to estimate CATE from both observational and RCT data. They combine predictions from base learners (regular machine learning methods) to account for the treatment assignment mechanism and confounding variables.

Variables

Before exploring the specific meta-learners, we first define the key variables used in our analysis:

- A is the treatment assignment indicator, where $A = 0$ is control - low PEEP regime, and $A = 1$ is treatment - high PEEP regime.
- X are the covariates of a patient.
- Y is the outcome, where $Y = 0$ is death, and $Y = 1$ is survival.

We have chosen the following meta learners to estimate the CATE:

S-learner

The S-learner uses a *single* model to estimate the CATE. In the S-learner approach, we concatenate the treatment indicator A with the features X and fit a single model to predict the outcome Y. The model learns the interaction between X and A, which allows for the estimation of the treatment effect. This follows the definition from Künzel et al.

$$\mu(x, a) := \mathbb{E}[Y|X = x, A = a] \tag{5}$$

To estimate the CATE, we compute the difference in predicted outcomes for treated ($A = 1$) and control ($A = 0$) groups:

$$\hat{\tau}_S(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0) \tag{6}$$

T-learner

The T-learner estimates the two response functions separately, therefore using *two* separate models for estimation. It involves training one model for the treated group and another for the control group. This also follows the definition from Künzel et al.

$$\begin{aligned} \hat{\mu}_1(x) &= \mathbb{E}[Y \mid X = x, A = 1] \\ \hat{\mu}_0(x) &= \mathbb{E}[Y \mid X = x, A = 0] \end{aligned} \tag{7}$$

The CATE is then estimated as the difference between these two models:

$$\hat{\tau}_T(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x) \tag{8}$$

DR-learner

The Doubly Robust (DR) learner is a CATE estimator that combines two approaches to reduce bias in causal inference. The DR-learner approach involves several steps to estimate the CATE. This follows the definition from Kennedy [4].

1. **Model the propensity score**: the probability a unit receives the treatment, defined as:

$$\pi(x) = \mathbb{P}(A = 1 \mid X = x) \tag{9}$$

2. **Estimate the response functions:**

Here the same approach is used as the T-learner (as shown in equation 7).

3. **Generate pseudo-outcomes**: combine the propensity score and response functions to create pseudo-outcomes. The pseudo-outcome is given by:

$$\widehat{\varphi}(Z) = \frac{A - \widehat{\pi}(X)}{\widehat{\pi}(X)\{1 - \widehat{\pi}(X)\}} \Big\{ Y - \widehat{\mu}_A(X) \Big\} + \widehat{\mu}_1(X) - \widehat{\mu}_0(X) \tag{10}$$

Here, $\widehat{\pi}(X)$ is the estimated propensity score, $\widehat{\mu}_1(X)$ and $\widehat{\mu}_0(X)$ are the estimated response functions for treated and control groups, respectively.

The final step involves using the pseudo-outcomes to fit a regression model, which estimates the CATE as:

$$\hat{\tau}_{DR}(x) = \mathbb{E}[\widehat{\varphi}(Z) \mid X = x] \tag{11}$$

The DR-learner offers several advantages thanks to its approach:

- **Reduced Bias with Two Models**: The DR-learner combines outcome regression and propensity score modeling, providing a doubly-robust property. This means that even if one of these models is misspecified, the estimator can still yield unbiased results.
- **Flexibility in Application**: The DR-learner allows for flexibility in choosing different algorithms for the outcome regression, propensity score modeling and pseudo-outcome regression. This flexibility can be beneficial for fine-tuning the model to fit specific data characteristics, as compared to the S- and T-learners that typically use a single model type.

The doubly-robust property has the potential to produce unbiased estimates despite unintended misspecifications, addressing the challenge of hidden confounders in observational datasets.

4 Experimental Setup and Results

4.1 Experiment Environment

The experiment was conducted using the Python programming language, more specifically using the Jupyter notebooks environment. For estimating the CATE, we leveraged the implemantion of the S-, T-, and DR- learners from the EconML library [10]. The code for this paper can be found at [11].

4.2 Data Simulation

Six simulations were conducted following the methodology described in the paper from Künzel et al. Additionally, two more simulations were performed for deeper insights:

- Simulations 1-6: As per Künzel et al., these simulations explore various conditions to evaluate the performance of meta-learners. The outcomes are continuous.
- Simulation 7: Replicates Simulation 1 but increases the propensity score to 12% to match the MIMIC-IV dataset's treatment rate. The outcomes are also continuous.

- Simulation 8: Combines the characteristics of Simulations 1 and 6 with a binary outcome to closely resemble the MIMIC-IV dataset.

The details of the simulations can be seen in Table 1.

All simulations had a maximum sample size of 10,000 across 10 iterations, constrained by computational limits. Gradient Boosting served as the base learner for S- and T-learners and the first stage of the DR-learner, while Linear Regression was used for the final stage of the DR-learner. Performance was evaluated using mean square error (MSE), leveraging the ability to observe both potential outcomes in simulations. The results can be seen in Figures 3 - 7.

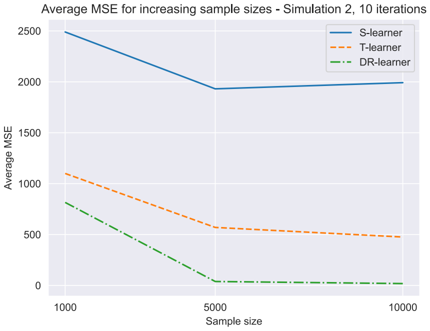


Figure 3: Simulation 2 (Complex Linear) - different linear response functions are applied across the feature space, with a propensity score of 0.5.

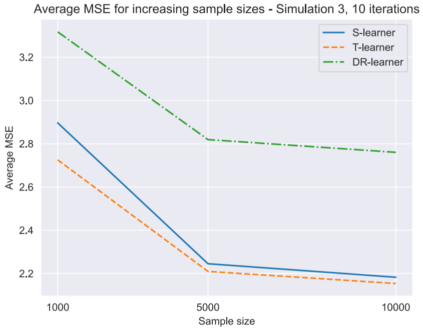


Figure 4: Simulation 3 (Complex Non-Linear) - non-linear response functions

Results

The S-learner performed well in most simulations, except when there were different linear response functions (Figure 3). This is likely due to its simplicity. The T-learner was consistently outperformed by both the S- and DR-learners in most simulations, except Simulation 3 (Figure 4). This is likely due to its approach of fitting separate models for the treated and control groups. The DR-learner excelled in specific scenarios where the dataset was unbalanced (Figure 6) and when confounding variables were present (Figures 5 and 7). This can be attributed to its design, allowing it to handle biases due to confounding and variance in treatment assignment more effectively. In simulations with different linear response functions applied across the feature space (Figure 3), the DR-learner performed exceptionally well. This suggests that the DR-learner is particularly adept at capturing complex linear interactions between covariates and treatment effects. However, in non-linear scenarios (Figure 4), the DR-learner did not perform as well, likely due to its reliance on Linear Regression in the final stage, which may not fully capture the

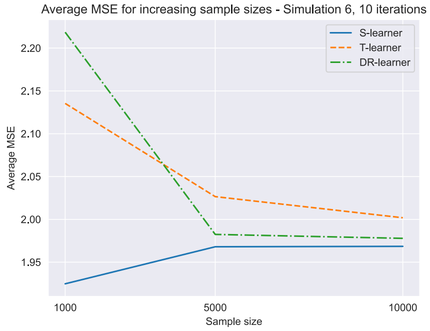


Figure 5: Simulation 6 (Beta Confounded) - Confounded data, using a beta distribution to simulate the propensity score.

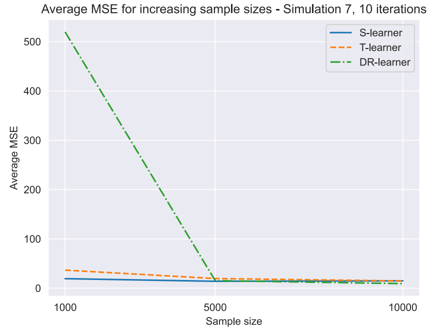


Figure 6: Simulation 7 - 12% of units receive treatment, with a simple CATE function to estimate.

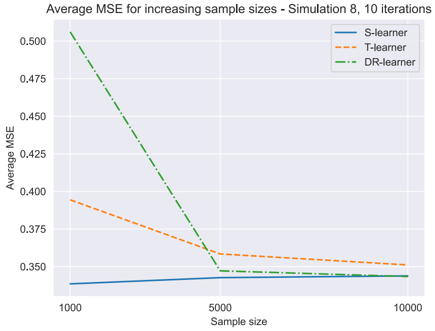


Figure 7: Simulation 8 (Beta Confounded) - Modification of Simulation 6 such that the response functions differ and are dependent on covariates, with binary outcomes.

Sim. no.	d	$e(X)$	$\mu_0(X)$	$\mu_1(X)$	Remarks
1	20	0.1	$X \cdot \beta + 5 \cdot \mathbb{I}(X_1 > 0.5)$	$\mu_0(X) + 8 \cdot \mathbb{I}(X_2 > 0.1)$	$\beta \sim U([-5, 5]^d)$
2	20	0.5	$X \cdot \beta_1$	$X \cdot \beta_2$	$\beta_1, \beta_2 \sim U([1, 30]^d)$
3	20	0.5	$\frac{1}{2} \cdot \varsigma(X_1) \cdot \varsigma(X_2)$	$-\frac{1}{2} \cdot \varsigma(X_1) \cdot \varsigma(X_2)$	$\varsigma(x) = \frac{2}{1+e^{-12(x-0.5)}}$
4	20	0.5	$X \cdot \beta$	$\mu_0(X)$	$\beta \sim U([1, 30]^d)$
5	20	0.5	$\begin{cases} X \cdot \beta_{1-2} & \text{if } x_{10} < -0.4 \\ X \cdot \beta_{3-6} & \text{if } -0.4 \leq x_{10} \leq 0.4 \\ X \cdot \beta_{7-9} & \text{if } 0.4 < x_{10} \end{cases}$	$\mu_0(X)$	$\beta_{k-l} = \begin{cases} \beta(i) & \text{if } k \leq i \leq l \\ 0 & \text{otherwise} \end{cases}$ $\beta \sim U([-15, 15]^d)$
6	20	$\frac{1}{4}(1 + \beta_{2,4}(X_1))$	$2 \cdot X_1 - 1$	$\mu_0(X)$	-
7	20	0.12	$X \cdot \beta + 5 \cdot \mathbb{I}(X_1 > 0.5)$	$\mu_0(X) + 8 \cdot \mathbb{I}(X_2 > 0.1)$	$\beta \sim U([-5, 5]^d)$
8	26	$\frac{1}{4}(1 + \beta_{2,4}(X_1))$	$2 \cdot X_1 - X_3 + 0.4 \cdot X_4 - X_5$	$\mu_0(X) - 0.1 \cdot X_9 + 0.9 \cdot X_{16} + 0.4 \cdot X_{20} - 0.2 \cdot X_{21} - 1$	Binary outcome

Table 1: Details of the simulations

non-linearity in the data. The Figures of the remaining simulations can be found in Appendix B.

4.3 MIMIC-IV

Further Pre-processing

To train the meta-learners, the MIMIC-IV dataset required further pre-processing. Initially, the dataset was scaled using the *MinMaxScaler* from the *scikit-learn* library. Approximately 7% of the data was missing. Therefore we compared two imputation methods:

- K-Nearest Neighbors
- Iterative imputer

To evaluate performance, we used the complete portion of the dataset, randomly removed 7% of the data, imputed the missing values, and compared the MSE of both imputers. Although both imputers performed well, as seen in Table 2, the iterative imputer introduced negative values, which would include data that could not normally be gathered. Therefore, we chose the K-Nearest Neighbors (KNN) imputer from the *scikit-learn* library with the parameters: *n_neighbors=11, weights='uniform'*.

Imputer	MSE
KNN Imputer	0.00137685
Iterative Imputer	0.00091196

Table 2: Imputer MSE comparison

After scaling and imputation, we will focus exclusively on the selected variables, as detailed at the end of Subsection 2.3: *Age, Weight, PF ratio, PO$_2$, Urea, pH, FiO$_2$, Plateau pressure and Driving pressure*.

Furthermore, the dataset includes the outcome variable *mort_28*, which indicates mortality after 28 days. For a more straightforward interpretation, we invert this outcome to focus on *survival*. Thus, a positive CATE would suggest that higher PEEP is beneficial for that specific patient, increasing their chances of survival.

Training the Meta-learners

In our research, we split the dataset into training and test sets, with a ratio of 80/20, to evaluate the performance of the chosen meta-learners.

S- and T-learners: We trained the S- and T-learners using Gradient Boosting and Linear Regression models. For Gradient Boosting, we set the parameters *n_estimators=100* and *random_state=768*.

DR-learner: For the DR-learner, we performed a grid search to select the best combination of propensity, response, and final models. The selection process involved evaluating the accuracy of treatment and outcome predictions and optimizing the AUQC.

We experimented with the following classifiers for the **propensity** and **response** models: *Logistic Regression, Gradient Boosting, XGBoost, SVM and K-Nearest Neighbors*. The propensity models were calibrated using *CalibratedClassifierCV* from *scikit-learn* with the parameters: *method='isotonic', cv=5*.

For the **final** model, we considered the regressors: *Gradient Boosting, XGBoost and Linear Regression*.

Based on the AUQC, we selected the best-performing models and parameters. Additionally, we evaluated a pure Gradient Boosting DR-Learner and a Logistic and Linear Regression DR-Learner.

Chosen Combinations for the DR-learner:

- SVM for Propensity and XGBoost for Response with XGBoost Regressor as the Final model
- SVM for Propensity and Logistic Regression for Response with XGBoost Regressor as the Final model
- SVM for Propensity and Gradient Boosting for Response with XGBoost Regressor as the Final model
- K-Nearest Neighbors for Propensity and Gradient Boosting for Response with Gradient Boosting Regressor as the Final model
- Gradient Boosting for all three models
- Logistic Regression for Propensity, Logistic Regression for Response, and Linear Regression as the Final model

The optimal parameters for the propensity and response models were determined as shown in Tables 3 and 4. This setup allowed us to systematically evaluate and select the optimal meta-learners for predicting treatment effects.

Propensity Models	Parameters
SVM	‘C=1.0‘ ‘kernel=’linear’‘
Logistic Regression	‘C=10.0‘ ‘solver=’lbfgs’‘
Gradient Boosting	‘learning_rate=0.1‘ ‘max_depth=3‘ ‘n_estimators=50‘
K-Nearest Neighbors	‘n_neighbors=7‘

Table 3: Optimal parameters for the propensity models

Response Models	Parameters
XGBoost	‘learning_rate=0.01‘ ‘max_depth=3‘ ‘n_estimators=50‘
Logistic Regression	‘C=1.0‘ ‘solver=’liblinear’‘
Gradient Boosting	‘learning_rate=0.01‘ ‘max_depth=3‘ ‘n_estimators=50‘

Table 4: Optimal parameters for the response models

We then trained the S-, T-, and DR-Learners with these models and parameters on 10 different train-test splits, calculating the average AUQC to assess their performance. The results can be seen in Table 5.

Results

The performance of the models varied significantly with different train-test splits; some splits yield better results, while others did not perform as expected. Averaging the AUQC and AUC across all 10 iterations suggests that Gradient Boosting and XGBoost are prone to overfitting. In contrast, linear models do not overfit but perform poorly overall, making it difficult to estimate the treatment effect reliably.

4.4 RCT Dataset Evaluation

The patient data from the RCT dataset includes a different set of features compared to the MIMIC-IV dataset. As a result, separate meta-learner models with fewer features were trained for this evaluation. The meta-learners intended for the RCT data were trained using the following features: *age, weight, pf_ratio, po2, ph, fio2, driving_pressure, plateau_pressure*. Note: the variable that is missing is *urea*.

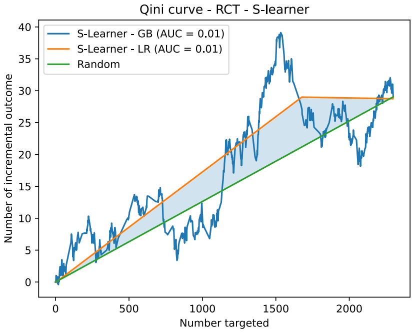


Figure 8: S-learner performance comparison with the largest area highlighted

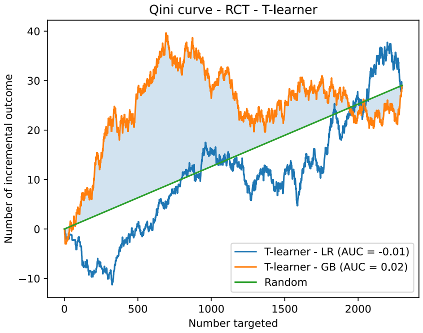
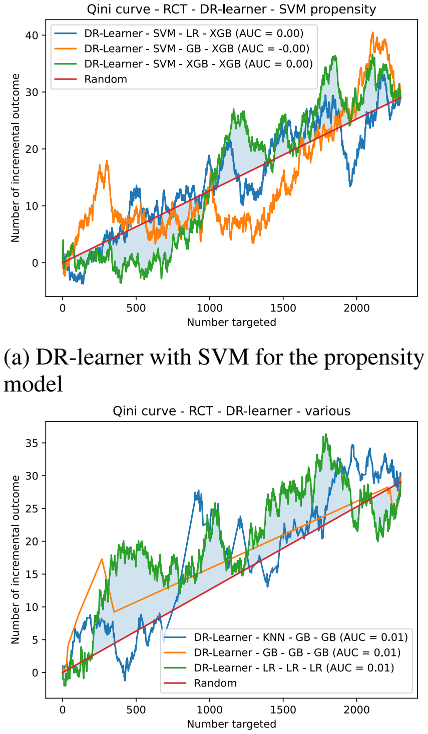


Figure 9: T-learner performance comparison with the largest area highlighted



(a) DR-learner with SVM for the propensity model

(b) The rest of the DR-learners

Figure 10: DR-learner performance comparison with the largest area highlighted

The figures and AUQCs indicate that the performance of the trained meta-learners is suboptimal. The results are summarized in the right-most column of Table 5 and in Figures 8, 9 and 10.

Results

The evaluation shows that most of the learners are unable to reliably predict the treatment effect. While the T-learner with gradient boosting shows some promise, the AUQCs are still close to zero. Overall, the experiment is inconclusive, indicating that verifying the treatment effect on the RCT data is unreliable.

Learner	Model	MIMIC-IV		RCT Dataset
		Train Set	Test Set	
S	Gradient Boosting	0.091742	-0.021610	0.005332
	Linear Regression	-0.004268	-0.00602	0.007350
T	Gradient Boosting	0.432664	0.009727	0.019314
	Linear Regression	0.048424	0.016777	-0.007351
DR	SVM Propensity and XGBoost Response with XGBoost Regressor Final	0.275965	0.016025	0.001503
	SVM Propensity and Logistic Regression Response with XGBoost Regressor Final	0.281178	0.010721	0.000868
	SVM Propensity and Gradient Boosting Response with XGBoost Regressor Final	0.281015	0.019345	-0.000503
	Logistic Regression Propensity and Logistic Regression Response with Linear Regression Final	0.022966	0.025802	0.009918
	Gradient Boosting Propensity and Gradient Boosting Response with Gradient Boosting Final	0.129593	0.018670	0.006207
	K-Nearest Neighbors Propensity and Gradient Boosting Response with Gradient Boosting Final	0.263782	0.042217	0.005886

Table 5: S-, T-, and DR-learners - Average AUQC of 10 tain-test splits of MIMIC-IV, and RCT AUQC Performance Comparison

5 Discussion

5.1 Result Analysis and Insights

Our results indicate that the meta-learners provided unreliable outcomes. Specifically, in the MIMIC-IV data, the performance was highly dependent on the train-test split, and averaging the results yielded a very low AUQC. Non-linear models also tended to overfit, further complicating the analysis.

Several factors could contribute to these suboptimal results. Firstly, the dataset size of 3941 samples may be insufficient for our analysis. Secondly, more time could have been spent on parameter analysis and model selection. Finally, we might have misidentified some confounding variables.

Furthermore, the evaluation using RCT data also showed inconclusive results. However, the consistent poor performance across all meta-learners suggests that our current data is inadequate for accurately estimating the CATE.

5.2 Limitations

There are several limitations to our research. Computational limitations may have constrained our ability to perform more extensive analyses. The size of the MIMIC-IV dataset is likely insufficient, as simulations indicate that larger sample sizes significantly improve performance. The chosen parameters might not have been optimal, potentially affecting results. Non-linear models showed tendencies to overfit.

5.3 Future Research

Future research should focus on several areas to improve the reliability of CATE estimation. Increasing the sample size is crucial for enhancing the reliability of treatment effect estimates. Efforts should be dedicated to fine-tune model parameters and explore more sophisticated methods such as Neural Networks, which were excluded due to their complexity and the time constraints of our research. Moreover, continuously reassessing the feature set to ensure that all relevant variables are included in the analysis is important. Finally, implementing regularization techniques to mitigate overfitting in non-linear models should be considered.

5.4 Alternatives

In addition to the meta-learners we have used, there are other methods available for estimating the CATE, such as the R-Learner [12] and the X-Learner [3]. The X-Learner, for example, performs well on unbalanced data, which is beneficial for the MIMIC-IV dataset where the number of patients receiving high PEEP is unbalanced. Other promising methods include TARNet [13], Causal Forests [14], and Multi-task Gaussian processes [15].

6 Responsible Research

In conducting our research, we have adhered to ethical standards and ensured the reproducibility of our methods. The MIMIC-IV dataset we utilized is already anonymized. The data has been de-identified, with all personal patient information removed or randomized. This ensures that the privacy of patients is protected and no sensitive information can be traced back to individuals.

We have taken several measures to handle the data responsibly. All data (except the RCT dataset) is stored locally on our systems, and all computations involving the data are performed locally as well. This minimizes the risk of data breaches and unauthorized access. Furthermore, we have made no attempts to re-identify any individuals within the dataset.

Additionally, we did not have direct access to the RCT dataset. This lack of access further ensures that we could not inadvertently compromise any additional data or violate any ethical guidelines associated with the use of RCT data.

It is also important to consider the potential implications of using these models in a clinical setting. Assigning treatments based on the predictions of machine learning models could lead to issues of fairness and equity. There is a risk that certain patient groups might be systematically favored or disadvantaged by the model's recommendations, leading to unequal access to treatment. To mitigate this risk, it is crucial to extensively evaluate the models for biases and ensure they are trained on diverse and representative data.

7 Conclusions

In this paper, we explored how the DR-learner, a machine learning-based method, can be used to predict survival outcomes in ICU patients under different PEEP regimes based on individual characteristics. Our main research question was to determine how the DR-learner compares to other CATE estimators, specifically the S- and T-learners, when evaluated on an RCT dataset.

Simulations showed that the S-learner performs well with non-linear response functions, while the DR-learner excels with unbalanced and confounded data, often matching or outperforming the S-learner. The T-learner was generally less effective.

However, when applying these methods to the MIMIC-IV observational dataset produced less promising results. Gradient Boosting and XGBoost models overfitted, while linear models, though not overfitting, performed poorly. Variability across train-test splits and low area under the Qini curve (AUQC) suggests that our models struggle to reliably estimate the treatment effect. This trend persisted even when evaluating the models on a smaller RCT dataset, where most learners failed to predict the treatment effect accurately, and the results remained inconclusive.

This research highlights the complexities of estimating CATE in ICU settings. Despite the theoretical robustness of certain meta-learners, practical application to real-world data remains challenging. The DR-learner's resilience to confounding and varying response functions in simulations underscores its potential in complex clinical data scenarios. However, the tendency of non-linear models to overfit indicates a need for cautious model selection and robust validation strategies in clinical applications.

In conclusion, while the DR-learner shows promise, further research is needed to refine these methods and improve their reliability in estimating treatment effects from real-world clinical data. Future work should focus on optimizing model parameters, exploring additional data sources, and mitigating overfitting.

References

[1] Allan J Walkey, Lorenzo Del Sorbo, Carol L Hodgson, Neill K J Adhikari, Hannah Wunsch, Maureen O Meade, Elizabeth Uleryk, Dean Hess, Daniel S Talmor, B Taylor Thompson, Roy G Brower, and Eddy Fan. Higher PEEP versus lower PEEP strategies for patients with acute respiratory distress syndrome. a systematic review and meta-analysis. *Ann. Am. Thorac. Soc.*, 14(Supplement_4):S297–S303, October 2017.

[2] Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-Wei H Lehman, Leo A Celi, and Roger G Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data*, 10(1):1, January 2023.

[3] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. U. S. A.*, 116(10):4156–4165, March 2019.

[4] Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. 2020.

[5] M Alan Brookhart, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *Am. J. Epidemiol.*, 163(12):1149–1156, June 2006.

[6] Matthias Briel, Maureen Meade, Alain Mercat, Roy G Brower, Daniel Talmor, Stephen D Walter, Arthur S Slutsky, Eleanor Pullenayegum, Qi Zhou, Deborah Cook, Laurent Brochard, Jean-Christophe M Richard, Francois Lamontagne, Neera Bhatnagar, Thomas E Stewart, and Gordon Guyatt. Higher vs lower positive end-expiratory pressure in patients with acute lung injury and acute respiratory distress syndrome: systematic review and meta-analysis. *JAMA*, 303(9):865–873, March 2010.

[7] Jesús Villar, Carmen Martín-Rodríguez, Ana M Domínguez-Berrot, Lorena Fernández, Carlos Ferrando, Juan A Soler, Ana M Díaz-Lamas, Elena González-Higueras, Leonor Nogales, Alfonso Ambrós, Demetrio Carriedo, Mónica Hernández, Domingo Martínez, Jesús Blanco, Javier Belda, Dácil Parrilla, Fernando Suárez-Sipmann, Concepción Tarancón, Juan M Mora-Ordoñez, Lluís Blanch, Lina Pérez-Méndez, Rosa L Fernández, and Robert M Kacmarek. A quantile analysis of plateau and driving pressures: Effects on mortality in patients with acute respiratory distress syndrome receiving lung-protective ventilation. *Crit. Care Med.*, 45(5):843–850, May 2017.

[8] Maureen O Meade, Deborah J Cook, Gordon H Guyatt, Arthur S Slutsky, Yaseen M Arabi, D James Cooper, Andrew R Davies, Lori E Hand, Qi Zhou, Lehana Thabane, Peggy Austin, Stephen Lapinsky, Alan Baxter, James Russell, Yoanna Skrobik, Juan J Ronco, Thomas E Stewart, and Lung Open Ventilation Study Investigators. Ventilation strategy using low tidal volumes, recruitment maneuvers, and high positive end-expiratory pressure for acute lung injury and acute respiratory distress syndrome: a randomized controlled trial. *JAMA*, 299(6):637–645, February 2008.

[9] Floris Devriendt, Jente Van Belle, Tias Guns, and Wouter Verbeke. Learning to rank for uplift modeling. *IEEE Trans. Knowl. Data Eng.*, 34(10):4888–4904, October 2022.

[10] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML, 2019. Version 0.x.

[11] Robert Melika. Ml-s-t-dr-cate-resp-care. https://github.com/RobertMelika/ML-S-T-DR-CATE-Resp-Care, 2024.

[12] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. 2017.

[13] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085. PMLR, 06–11 Aug 2017.

[14] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.*, 113(523):1228–1242, July 2018.

[15] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. 2017.

A Feature Selection

Feature importance when estimating outcome and treatment assignment when the *peep* feature is included.

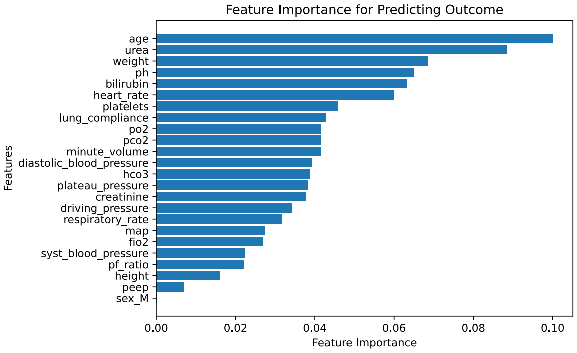


Figure 11: Feature importance when estimating outcome with the *peep* feature included

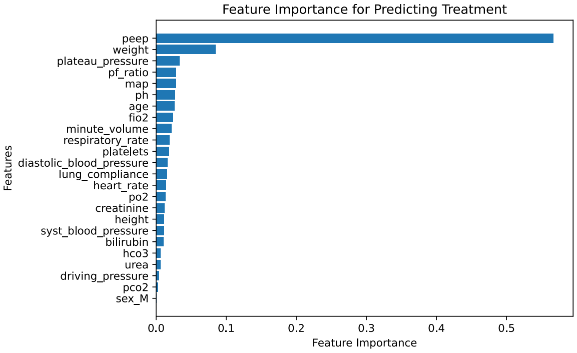


Figure 12: Feature importance when estimating treatment assignment with the *peep* feature included

B Simulation Figures

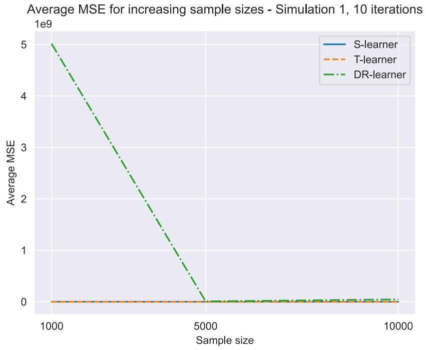


Figure 13: Simulation 1 - only a small percentage of units receive treatment (1%), with a simple CATE function to estimate.

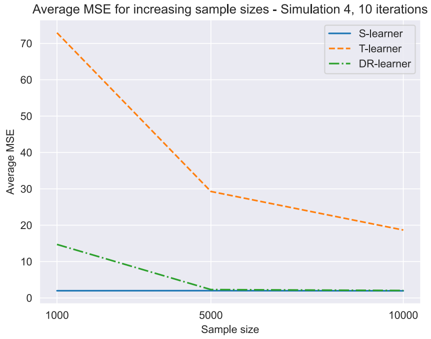


Figure 14: Simulation 4 (Global Linear): Both response functions are globally linear in this scenario, resulting in a zero treatment effect.

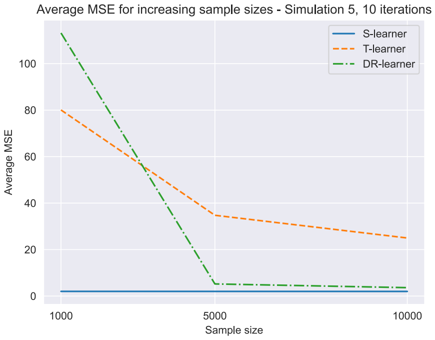


Figure 15: Simulation 5 (Piecewise Linear) - feature space divided into three segments, each with a distinct linear response function