Optimizing Mechanical Ventilation Support for Patients in Intensive Care Units

An Analysis of Deep Learning Methods for Personalizing Positive End-Expiratory Pressure Regime

Mircea-Petru Anica-Popa[2]

Supervisor(s): 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, 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: Mircea-Petru Anica-Popa
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Rickard Karlsson, Jim Smit, Jasmijn Baaijens

Abstract

In the intensive care unit (ICU), optimizing mechanical ventilation settings, particularly the positive end-expiratory pressure (PEEP), is crucial for patient survival. This paper investigates the application of neural network-based machine learning methods to personalize PEEP settings in the ICU, aiming to improve patient survival outcomes. The research focuses on two specific algorithms, TARNet and CFR, evaluating their ability to estimate individualized treatment effects of lower versus higher PEEP regimes. The study is structured into three phases: controlled simulations, application to the MIMIC-IV dataset, and validation using a randomized control trial dataset. TARNet and CFR showed potential for estimating the individualized treatment effects but required large datasets for optimal performance. In the case where limited data is available, these models are upstaged by simpler learners, such as the S- and T-learners. The study concludes that while neural network-based methods hold promise for personalizing ICU treatment, their efficacy is heavily influenced by data availability and quality.

1 Introduction

Mechanical ventilation is a critical supportive therapy for patients with acute respiratory failure admitted to the intensive care unit (ICU) [1]. A key setting in mechanical ventilation is the positive end-expiratory pressure (PEEP). High PEEP may reduce lung stress and strain, but it also has potentially harmful side effects, such as ventilation-induced lung injury [2]. These risks complicate the decision of medical professionals on how to set the PEEP in mechanical ventilation.

The question of whether high or low PEEP should be used for mechanical ventilation in ICU patients remains controversial despite numerous randomized controlled trials comparing low versus high PEEP regimes [3–6]. The research hypothesis is that some patients benefit more from high PEEP regimes than others, depending on certain patient characteristics. For instance, one study found that high PEEP was associated with improved survival rates for patients suffering from acute respiratory distress syndrome [7]. Moreover, another study discovered that patients with hyperinflammatory responses have a better chance of survival when treated with low PEEP, while individuals with less severe inflammatory reactions have a higher likelihood of survival when treated with high PEEP [8]. If there existed a method that could infer what regime is more suitable based on the characteristics of an individual, treatment strategies could be personalized accordingly.

The characteristics of this decision could make it suitable to be modeled as a causal inference task of estimating individualized treatment effects (ITE) [9]. *Causal inference* refers to the process of determining and estimating the impact of certain actions on outcomes in a population, also known as the *treatment effect* [10]. In this case, this would be the effect on survival outcomes of ICU patients from following either a low or high PEEP regime. Multiple methods based on different types of machine learning algorithms have shown promising results in estimating ITE and could potentially be applied to the problem of determining the optimal PEEP regime for individual ICU patients [11–16].

The main research question this paper aims to answer is: *"How can neural network-based machine learning methods such as TARNet and CFR be used to personalize treatment strategies in the ICU by estimating the individualized treatment effects of lower versus higher PEEP regimes for mechanical ventilation on patient outcomes?"*. The paper focuses on evaluating machine learning methods based on neural networks, specifically TARNet and CFR [13], by comparing them with two simpler baseline metalearners, respectively, the S- and T-learners [11], throughout multiple experiments. The methodological approach included firstly running the algorithms on various instances of simulated data covering a wide range of possible scenarios. Secondly, they were applied to the MIMIC-IV, a publicly available ICU database intended to support research studies and educational material [17], to better understand how these methods would behave in a real-world scenario. Thirdly, the models were evaluated using randomized control trials to provide a form of external validation for their performance. By analyzing these experiments and interpreting their results, this paper is expected to contribute to filling the knowledge gap by gaining insights into how these methods can be applied to real-world ICU data. Understanding how the results can be interpreted in a clinical context could potentially lead to improved patient care by personalizing treatment strategies in the ICU.

The main research question can be broken down into the following sub-questions, which will be used to structure the research process and the resulting paper:

1. What are the potential benefits and downsides of using the chosen methods for estimating individualized treatment effects on ICU patient outcomes?
2. How can the results of applying the chosen ML algorithms to the MIMIC-IV dataset for the individual treatment effect estimation be interpreted in the context of ICU patient care?

The paper is structured as follows. Section 2 defines the relevant terms and background knowledge. Section 3 describes the research methodology used during the project and introduces the main ITE estimators evaluated during the experiments. Section 4 elaborates on the setup and outcomes of the experiments. Section 5 discusses the obtained results, the potential advantages and drawbacks of the evaluated methods, and the limitations encountered throughout the project. Section 6 reflects on the ethical implications of the project and examines the reproducibility of the results. Finally, Section 7 draws the main conclusions and provides recommendations for future research.

2 Problem Setup

Causal inference is a methodological approach that allows researchers to determine causal relationships by examining the conditions under which an effect occurs [10, 18]. This approach differs fundamentally from classical machine learning, which is often geared towards prediction. The focus of causal inference is to comprehend the consequences of interventions or actions, a critical distinction in domains such as healthcare [15]. In those settings, it is paramount to understand the effect of a treatment on a patient rather than merely predicting the health status of a patient in the absence of any intervention.

The focus of this project is on estimating the individualized treatment effect of lower versus higher PEEP regimes for mechanical ventilation on patient outcomes. The ITE is the effect a specific treatment has on a particular individual compared to other possible treatments. Formally, the ITE for a specific individual can be defined as follows, where $Y(1)$ and $Y(0)$ represent potential outcomes under the treatment and control, respectively, and $X = x$ denotes the features of the individual:

$$ITE(x) = E[Y(1) - Y(0)|X = x] \tag{1}$$

The main difficulty of using machine learning techniques to estimate this function is that, for any given individual, only the outcome under the assigned treatment condition can be observed, and its counterfactual remains unknown. This is known as the *fundamental problem of causal inference* and is the reason why classical machine learning methods cannot directly be trained on the difference $Y(1) - Y(0)$ [18].

Another challenge in causal inference is that, in observational studies, the assignment of treatments is not random, which can lead to *confounding*. Confounding occurs when the treatment assignment is correlated with other observed or unobserved covariates that also affect the outcome [10]. This correlation can introduce bias in the estimation of the treatment effect because it is unclear whether the outcome is due to the treatment or the confounding variables. Unlike observational studies, randomized controlled trials (RCTs) assign treatments at random, which eliminates the correlation between the treatment and other covariates, thus preventing confounding [10, 18].

Propensity scores can be employed to address confounding in observational studies. The propensity score of an individual represents the probability of receiving the treatment given a set of observed covariates (the features of an individual) [19]. By matching or weighting on propensity scores, researchers can balance the distribution of observed covariates between the treated and control groups, thereby reducing the bias due to confounding [11, 12, 16]. This method attempts to mimic the conditions of an RCT within the confines of observational data, thus allowing for more reliable causal inferences. Formally, the propensity score can be defined as follows, where $T = 1$ indicates the treatment group and $X = x$ represents the observed covariates for the individual:

$$e(X) = P(T = 1|X = x) \tag{2}$$

To be able to use machine learning methods to estimate the ITE, three assumptions need to be made [13, 14, 20, 21]. The first assumption is *consistency*, which requires that the potential outcome of an individual under a particular treatment is the same as the observed outcome if the individual receives that treatment. This condition can be formally described as $Y = Y(T)$. The second assumption is *unconfoundedness* (or *ignorability*), which implies that the treatment assignment is independent of the potential outcomes given a set of observed covariates. This property can be formally written as $Y(0), Y(1) \perp T|X = x$. The third assumption is *overlap*, which states that every individual must have a positive probability of receiving either treatment. This requirement can be formally defined as $0 < P(T = 1|X = x) < 1, \forall x \in X$. These conditions allow researchers to treat observational data as having come from a conditionally randomized experiment [10].

3 Methodology

3.1 MIMIC-IV dataset

The MIMIC-IV dataset is a collection of deidentified health-related data intended for a wide array of applications within healthcare. It is the successor to the MIMIC-III dataset, incorporating contemporary data and improving on various aspects of its predecessor. The dataset was collected at the Beth Israel Deaconess Medical Center and includes patient measurements, orders, diagnoses, procedures, treatments, and deidentified free-text clinical notes [17].

The dataset that will be referenced throughout this paper contains data on 3941 individuals, each with 23 covariates — one of which is categorical, and the rest are continuous. The treatment variable in this dataset is the PEEP regime assigned to each patient, categorized as low or high, and the outcome variable is mortality after 28 days. It is also worth noting that the data is unbalanced, with only 12% of individuals having high PEEP assigned as treatment.

Several preprocessing steps were taken to prepare the dataset for machine learning. Firstly, in order to apply algorithms using neural networks to the data, no features should be categorical. Therefore, the categorical feature “sex” has been numerically encoded, with “M” represented as 0.0 and “F” as 1.0. Additionally, the boolean treatment and outcome variables have been converted to numerical form, with “False” represented as 0.0 and “True” as 1.0. Secondly, the data has been normalized across features using min-max normalization to ensure uniformity in scale.

Another crucial step involved addressing missing values, as around 7% of the data was papered as absent. To tackle this issue, two imputation methods from the *scikit-learn* library [22] were employed. The first one, the KNNImputer, leverages the k-Nearest Neighbors algorithm, identifying the k closest data points with similar features to the missing value and computing the mean to estimate the missing value. This method is particularly reliable as it does not rely on assumptions about the data distribution. On the other hand, the IterativeImputer applies a more complex strategy, iteratively modeling each feature with missing values as a function of other features in a regression framework. One downside of this method is that it can sometimes lead to the generation of unrealistic values. After evaluating the performance by introducing artificial missing data and calculating the Mean Squared Error (MSE) between the imputed and original values, it was observed that while the IterativeImputer provided a slightly more accurate imputation, it was prone to producing negative values for inherently positive attributes. As a result, the decision was made to use the data imputed with KNNImputer due to its reliability.

The final step was selecting the relevant covariates to be used for the machine learning task. To reduce bias, the models will only be trained using features that are considered possible confounders or have a strong influence on the outcome. The set of potential confounders includes the *PaO_2/FiO_2 ratio* and its components (*PaO_2* and *FiO_2*), as well as the *plateau pressure* of a patient, as evidenced by a study from 2010 [7]. Additionally, a study from 2021 has shown that demographic factors such as *age* and *sex* significantly impact both treatment assignment and mortality [23], making them potential confounders as well. Other factors to consider are blood oxygenation metrics such as *$PaCO_2$* and *driving pressure* and inflammation markers like *bilirubin*, *platelets*, and *urea* [8].

To validate the list of possible confounders, it was attempted to find an association, and thus possible causation, between the features present in the MIMIC-IV dataset and the outcome variable by selecting features that highly affect the outcomes of the patients. In our specific scenario, empirical evidence suggests that *sex* and *platelets* are not strongly associated with the outcome. However, it found that *weight* has a significant impact on mortality. As a result, it was decided that *weight* would be included in the training features in addition to the already presented covariates. At the same time, *sex* and *platelets* would be removed from the analysis.

3.2 Estimators

This paper focuses on ITE estimators that are based on neural networks. Neural networks are suitable for this task because they are able to model complex, non-linear relationships and capture high-dimensional interactions in the data. However, in certain scenarios, they may not be the best choice. For example, they tend to overfit the data when dealing with a small sample size. Additionally, if transparent decision-making processes are necessary, their "black box" nature may not be suitable. Nevertheless, neural networks have been successful in learning ITEs and estimating counterfactual outcomes, which are the outcomes that would have occurred under a different treatment condition [13, 20, 24, 25]. For the scope of this project, the following estimators were considered: the S-learner, the T-learner, TARNet, and CFR.

The S-learner [11] is a meta-learner that trains a single machine-learning model on all the available training data, regardless of treatment assignment. The model $\hat{\mu}$ is trained on the outcome Y as a function of both the covariates X and the treatment indicator T, as shown in the equation $\hat{\mu} = M(Y \sim (X, T))$. Once trained, the S-learner predicts the outcome for an individual as if they had received the treatment ($\hat{\mu}(x, 1)$) and as if they had not ($\hat{\mu}(x, 0)$). The ITE for an individual x is then the difference between these two predicted outcomes, calculated as $\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$. The advantages of the S-learner include its simplicity and the utilization of all available data during training. However, since the treatment is not given a special role, the base model of the learner can choose to ignore it, which can bias the ITE towards 0.

Conversely, the T-learner [11] is a meta-learner that trains separate machine-learning models for the treatment and control groups. For the control group, the model $\hat{\mu}_0$ is trained on the control outcomes Y^0 as a function of the covariates of individuals under control X^0, resulting in the estimation $\hat{\mu}_0 = M_0(Y^0 \sim X^0)$. Similarly, for the treatment group, the model $\hat{\mu}_1$ is trained, yielding $\hat{\mu}_1 = M_1(Y^1 \sim X^1)$. The ITE for an individual is then estimated by taking the difference between the potential outcome under treatment, predicted by $\hat{\mu}_1(x)$, and its counterfactual, predicted by $\hat{\mu}_0(x)$, which is expressed as $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$. The T-learner can capture different relationships in the treatment and control groups, allowing it to model varying treatment effects. Thus, the T-learner is expected to perform well when the treatment effect varies with the covariates. However, it uses only a fraction of the data to train each model separately, which could be a significant drawback if the sample size is small.

TARNet (Treatment Agnostic Representation Network) and CFR (Counterfactual Regression) [13] are two variants of an estimator that uses deep neural networks to estimate the causal effect of a treatment on the outcome of an individual. This configuration allows the estimator to learn complex nonlinear relationships flexibly.

TARNet works by training one neural network to learn a shared representation of the covariates independent of the treatment assignment. Additionally, to avoid losing the influence of the treatment during training, the estimator uses the learned representation to train separate "heads" of a second neural network, the hypothesis network, for the treatment and control group. It should be noted that only the head corresponding to the observed treatment is updated with each training sample. This approach allows for leveraging the statistical power of the whole dataset in the representation network while maintaining the effect of the treatment assignment in the two independent heads. The ITE for a new individual is estimated by taking the difference between the two potential outcomes predicted by each of the heads of the neural network after passing its covariates as input.

While TARNet only tries to accurately predict the potential outcomes, CFR extends the idea by adding a balance regularization term to the loss function of the algorithm. This term includes an Integral Probability Metric, denoted as IPM_G, which measures the discrepancy between the treated and control group distributions in the shared representation space. The "G" refers to the specific metric used, such as the Wasserstein distance or the Maximum Mean Discrepancy (MMD), to penalize the model if the distributions $P(x|t = 1)$ and $P(x|t = 0)$ diverge. The intuition behind this is that the difference between the treated and control distributions is 0 in a randomized controlled trial. Therefore, the balance regularization term will also be 0. Thus, making the treated and control distributions similar in the shared representation space can help to adjust for imbalanced data and improve the overall performance of the causal effect estimates. A graphical representation of the architecture of the algorithm can be seen in Figure 1.

The strength of the architecture of the algorithm, a shared representation network, and separate heads for each treatment assignment, paired with the balance regularization term,

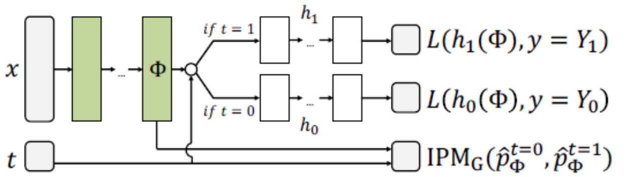


Figure 1: Neural network architecture for TARNet / CFR. L is a loss function, $\mathrm{IPM_G}$ is the metric used for balance regularization. The image was taken from [13].

has been demonstrated through both theoretical and empirical work [13]. Firstly, these methods are based on a new theoretical framework that limits the expected estimation error of ITE by combining the standard generalization error with the distributional distance within the representation. Secondly, empirical validations on real and simulated datasets affirm that TARNet and CFR often surpass both simpler, traditional models, such as Ordinary Least Squares and k-Nearest Neighbors, and more complex models, like Bayesian Additive Regression Trees and Causal Forests, achieving lower error rates in ITE predictions.

3.3 Experimental approach

The methodology adopted for this research project begins with a comprehensive literature study. This involves reviewing existing studies in the field, understanding the current state of knowledge, and identifying gaps that this project aims to address. The next step is the experiment setup, which involves generating synthetic data resulting in several simulations. These simulations are used to validate the results of the estimators. Additionally, the MIMIC-IV dataset is preprocessed to convert categorical features to numerical representations and impute missing values, preparing the data for model training.

The core of the methodology is the training and evaluation of the machine learning models. This is done both on the simulated data and the MIMIC-IV dataset. The training, evaluation, and analysis process is not a one-time process but an iterative one. The models are continually trained and evaluated to extract the best possible hyperparameter configuration.

The models are first evaluated in six different scenarios to analyze their performance in various circumstances. This helps validate the results observed on the MIMIC-IV dataset and establishes reasonable performance expectations for other real-world datasets. The data generation process is initiated by specifying the number of features d, the propensity score $e(x)$, and the response functions $\mu_0(x)$ and $\mu_1(x)$. Next, the feature vector $X_i \sim \mathcal{N}(0, I_d)$ is simulated, where I_d is the d-dimensional identity matrix. The potential outcomes $Y_0 = \mu_0(x) + \epsilon_0$ and $Y_1 = \mu_1(x) + \epsilon_1$ are then calculated, where $\epsilon_0, \epsilon_1 \sim \mathcal{N}(0, 1)$ represents random noise. The treatment assignment for each sample $T_i \sim \mathrm{Bern}(e(X_i))$ is simulated as a Bernoulli random variable. Finally, the observed outcome Y_i is set to either Y_0 or Y_1 depending on the treatment assignment T_i. Additionally, $ITE = Y_1 - Y_0$ is appended to each sample, resulting in a sample of the form (X_i, T_i, Y_i, ITE_i).

The algorithms are applied to samples of varying sizes, ranging from 3000 to 15000. This range was selected to keep the simulations similar to the MIMIC-IV dataset regarding available training data while also observing the behavior of the models on a larger number of samples. Since both potential outcomes are produced during the simulation study, the Mean Squared Error (MSE) between the actual ITE and the estimated ITE can be used to assess the models. To ensure the reliability of the results, each experiment was repeated 50 times, and the averages were presented visually using plots.

Secondly, the models are applied to the MIMIC-IV dataset. Since it comprises real-world data, only the outcome under the assigned treatment is observed. Thus, it is impossible to calculate the MSE between the actual ITE and the estimated ITE. Instead, the Qini curve [26] was chosen, which is a metric used to evaluate the performance of a model by measuring the incremental impact of a treatment and comparing it to a random selection. The curve helps identify the most responsive individuals or units to the treatment. In addition, to complement the visual representation of the Qini curve, the area under the curve (AUC) was analyzed. It represents a numerical metric extracted from the Qini curve that allows for easy comparisons between models. A good model would result in a curve that significantly diverges from the curve of the random pick, thus having a high AUC. These two metrics were used during the hyperparameter tuning grid search to determine the model with the most promising results. Each experiment was repeated 200 times, and the averages and standard deviations were recorded.

Finally, the estimators are evaluated using a dataset comprised of randomized control trials. This provides a benchmark for the performance of the models considered in this project by externally validating their effectiveness. The iterative process of training and refinement, coupled with thorough validation and comparison with other strong estimators, ensures the reliability and validity of the research outcomes.

4 Experiments and Results

For the implementation of the S-learner and T-learner, TensorFlow 2.10 [27] was used, with both learners utilizing deep neural networks as base models. The TARNet and CFR models were implemented using the *catenets* package [28]. This package provides a specialized framework for causal inference with neural networks. It is particularly well-suited for the tasks at hand due to its focus on estimating causal effects using deep learning techniques. For the IPM term used by CFR to penalize imbalance, the Maximum Mean Discrepancy distance was utilized.

All models were assigned ELU [29] as the activation function for all neurons in the hidden layers. The output layer used the sigmoid activation function for binary outcomes, while no activation function was employed for continuous cases. The Adam optimizer [30] was applied to all models, with an initial learning rate of 0.001. The chosen loss function was the Mean Squared Error for continuous outcomes, appropriate for the simulations, and Binary Cross-Entropy loss for binary outcomes, suitable for the MIMIC-IV dataset. These loss functions are standard choices for regression and classification tasks, respectively.

The hyperparameters for the models were chosen through a grid search on the MIMIC-IV dataset. Grid search was used to thoroughly explore various combinations of parameter settings to find the most effective and robust ones. Due to computational constraints, it was not feasible to perform a grid search for each simulation. Therefore, the simulations were conducted using the model resulting from the MIMIC-IV grid search. As a result of the grid search, the neural networks used as base models for both metalearners were comprised of three hidden layers of 200 neurons, followed by another hidden layer of 100 neurons. TARNet and CFR both used two fully connected neural networks, one representation network with three hidden layers of 200 neurons and one hypothesis network with one layer of 100 neurons. Additionally, all models employed a batch size of 200 samples.

The data was divided for training, validation, and testing using a 64/16/20 ratio for all experiments. All models were trained for 1000 epochs with an early stopping criterion based on the validation loss. The training process was stopped if no improvement in the validation loss was observed for 10 consecutive epochs. This approach helps prevent overfitting and ensures that the models generalize well to unseen data.

4.1 Simulation experiments

Six simulation experiments were conducted to evaluate ITE estimators under different conditions. The data for these simulations was generated following the methods outlined in section 3 of our paper. All simulations involved binary treatment and continuous outcomes, but they varied in terms of treatment balance, confounding variables, and the complexity of the ITE.

The simulations ranged from scenarios with an unbalanced treatment distribution, where only 10% of subjects were treated, having a simple ITE, to balanced treatment conditions with 50% of subjects treated. The balanced scenarios varied in complexity from simple linear ITEs to complex nonlinear ITEs and even included scenarios where the ITE was zero, both with and without confounding variables. The parameters for each simulation are documented in Appendix A. The results of the experiments were plotted in Figure 2.

A baseline comparison was made against a predictor that assigns the Average Treatment Effect (ATE) as the ITE for each patient. The results, plotted for clarity, revealed that none of the models exhibited signs of overfitting, and all showed significant improvement over the baseline ATE predictor for the simulations with ITE $\neq 0$. For the simulation with no treatment effect, the ATE estimator has the lowest MSE, with the S-learner following very closely.

Analyzing the averaged MSE for the simulations illustrated in Figure 2, the S-learner showed consistent and low MSE in both training and testing, indicating stable performance and good generalization from training to unseen data. The T-learner recorded higher variability in MSE, particularly in Figures 2a and 2e, suggesting it may be more sensitive to the specific simulation conditions. TARNet and CFR produced very similar results across all simulations. Their performance was lower than that of the S-learner but comparable to that of the T-learner, yielding smaller errors in simulations 2a, 2c, 2d, and 2e. These results indicate that TARNet and CFR might perform well in practice with proper regularization or tuning.

4.2 MIMIC-IV experiments

In the initial phase of our experiments on the MIMIC-IV dataset, a grid search was conducted specifically for the TARNet model to optimize its configuration settings. This involved assessing the average area under the cumulative gain curve over 200 runs, with the aim of determining the most effective model parameters. The parameters considered in this grid search were the following:

- Representation layers: { 1, 2, 3 }
- Hypothesis layers: { 1, 2, 3 }
- Neurons per representation layer: { 20, 50, 100, 200 }
- Neurons per hypothesis layer: { 20, 50, 100, 200 }
- Batch sizes: { 100, 200, 500, 700 }

Following this stage, the top 60 configurations, selected based on the mean AUC, were subjected to further hyperparameter tuning. The second round of optimization was dedicated to the CFR model, where the grid search was expanded to include a range of regularization strengths, denoted by alpha. The values for the alpha parameter were selected according to the formula $\{10^{k/2} | k \in \{-6, -4, -2, 0, 2, 4, 6\}\}$.

The grid search results revealed that certain hyperparameter configurations led to models achieving a positive mean AUC, indicating improvement over random assignment. The scores of the top five models can be found in Table 1.

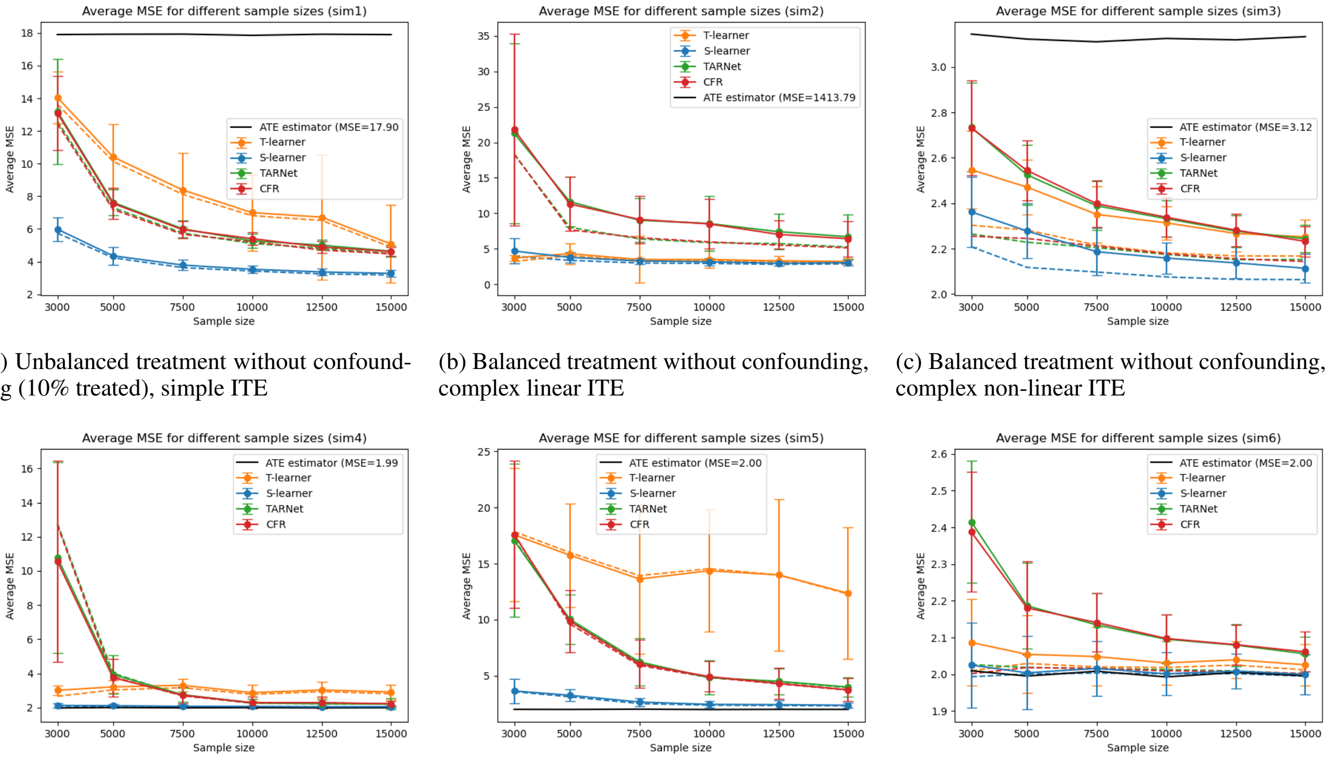
Table 1: Top scoring hyperparameter configurations of the grid search in terms of mean AUC (rounded to the nearest hundredth). (A, B, C, D, E, F) represents A representation layers with B neurons, C hypothesis layers with D neurons, batch size E, alpha F.

Configuration	Mean	Std
(3, 200, 1, 100, 200, 0.001)	0.69	1.62
(2, 200, 1, 20, 200, 0.001)	0.65	1.71
(3, 200, 1, 200, 200, 0.001)	0.64	1.83
(2, 200, 1, 20, 200, 0)	0.61	1.93
(3, 200, 1, 50, 200, 0.001)	0.61	1.70

Despite using the best scoring model configurations, the Qini curve exhibited a significant degree of randomness, notably changing with each retraining and evaluation of the model. This variability was likely due to the limited amount of training data, which made the performance of the model heavily dependent on the initial weight initialization of the neural networks. Table 2 shows the mean, standard deviation, and maximum AUC for each of the models applied to the MIMIC-IV dataset. It can be observed the simpler models, the S- and T-learners, greatly outperform TARNet and CFR in terms of mean area, but also exhibit a higher variance.

4.3 RCT Validation Experiments

The validation of the machine learning models was further extended through the use of three randomized control trial



(a) Unbalanced treatment without confounding (10% treated), simple ITE

(b) Balanced treatment without confounding, complex linear ITE

(c) Balanced treatment without confounding, complex non-linear ITE

(d) Balanced treatment without confounding, global linear response, ITE = 0

(e) Balanced treatment without confounding, piecewise linear response, ITE = 0

(f) Balanced treatment with measured confounding, ITE = 0

Figure 2: Mean Squared Error of the estimators across the simulations averaged over 50 runs. The dashed line shows MSE on the train set, the full line shows MSE on the test set, and the error bars show one standard deviation. The ATE for simulation 2 was too high to show on the plot but can be found in the legend.

Table 2: The mean, standard deviation, and maximum AUC for each model on the MIMIC-IV dataset (rounded to the nearest hundredth).

Model	Mean	Std	Max
S-learner	3.33	2.28	9.41
T-learner	3.26	2.05	10.02
TARNet	0.31	1.78	5.75
CFR	0.69	1.62	5.28

(RCT) studies [4–6], having a combined total of 2299 individual samples, with 549, 983, and 767 samples respectively. The dataset compiled from these studies was balanced, with 49.4% of the individuals being assigned to high PEEP.

To maintain consistency with the previous experiments, the same imputation method was applied to the MIMIC-IV dataset to handle missing data within the RCT dataset. Moreover, the same set of confounders was employed to adjust for potential biases in the analysis, with the exception of *bilirubin* and *urea*, which were not included in the feature set of the RCT studies. Given the high variance observed in the performance of the models and the constraint that the experiment could only be conducted once, it was crucial to select the most promising instantiation for each model. This selection process involved training and evaluating 200 instances of each model and choosing the one with the highest AUC for evaluation in the RCT experiments.

Drawing from the insights gained during the MIMIC-IV experiments, it was expected that the S-learner and the T-learner would significantly outperform TARNet and CFR in terms of AUC. However, in RCT dataset experiments shown in Table 3, the T-learner emerged with the highest mean AUC, indicating its effectiveness in randomized treatment assignments. The S-learner showed reasonable performance, though not as dominant as in the MIMIC-IV dataset. TARNet displayed moderate performance, better than both the S-learner and CFR, while CFR's notably low mean AUC was unexpected, highlighting that its balance regularization may not be as beneficial in RCTs as in observational studies. All models experience a high degree of variability, as shown by the wide confidence intervals.

5 Discussion

5.1 Results from simulation experiments

For the simulation experiments, the consistent performance of the S-learner, particularly in simulations without any treat-

Table 3: AUC with 95% confidence intervals for each model on the RCT dataset (rounded to the nearest hundredth).

Model	Area	95% CI
S-learner	0.22	[-2.31, 2.71]
T-learner	0.58	[-1.89, 2.97]
TARNet	0.40	[-2.16, 2.91]
CFR	0.04	[-2.40, 2.43]

ment effect, suggests that it was able to effectively model the ITE for the entire sample. This implies that in most scenarios, the S-learner's approach of training a single model on all data, without differentiating between treatment groups, is sufficient and effective, as expected [11]. The T-learner's underperformance compared with the other models in the unbalanced simulation, as shown in Figure 2a, could be due to its separate modeling of treatment and control groups. While this approach is theoretically advantageous for capturing different treatment effects, it may suffer from data sparsity in practice, particularly in unbalanced datasets where one group has significantly fewer samples [11].

TARNet and CFR were outperformed by the S-learner across all simulations, illustrated in Figure 2. They were also surpassed by the T-learner in the simulations displayed in Figures 2b and 2f. However, they managed to achieve better results in the simulations shown in Figures 2a, 2c, 2d, and 2e, likely because they combine the strengths of shared representation learning with treatment-specific modeling. This dual approach allows them to leverage the entire dataset for learning a common representation while also capturing treatment-specific nuances. However, it can be observed that while TARNet and CFR are powerful and can adapt to multiple scenarios, their performance relies heavily on the availability of sufficient data to learn from, which can be a limiting factor in real-world scenarios where data may be scarce or expensive to obtain. This data-hungry behavior can be most clearly observed in the simulation shown in Figure 2f as the models required between 7500 and 10000 samples to achieve a performance comparable to the two metalearners. It is also worth noting that TARNet and CFR have very similar scores across all simulations, likely due to the small alpha value.

5.2 Results from MIMIC-IV experiments

In the context of the MIMIC-IV dataset, the TARNet and CFR algorithms showed a significant difference in their mean AUC scores (0.31 versus 0.69 - Table 2). TARNet's mean AUC was notably lower than CFR's, suggesting that the balance regularization term included in CFR might provide a significant advantage in this unbalanced dataset. The standard deviation and maximum AUC values indicated high variance in TARNet's performance. In contrast, CFR's slightly lower standard deviation and maximum AUC suggested a more consistent performance, albeit still with considerable variance. This behavior likely resulted from the limited number of samples available in the MIMIC-IV dataset. The simulations showed that TARNet and CFR typically required a larger sample size (between 7500 and 10000) to produce reliable results. Given the constraints of the MIMIC-IV dataset, the two models exhibited an over-reliance on the specific data splits and the initialization of the weights of the neural networks, leading to inconsistent performance across different runs.

Furthermore, TARNet and CFR were originally designed to work with continuous outcomes and mean squared error as a loss function [13]. The shift to a binary outcome with binary cross-entropy loss in the MIMIC-IV dataset represents a change in the conditions for which these algorithms were optimized. This mismatch between the theoretical background of TARNet and CFR and the actual data scenario in MIMIC-IV could result in unexpected behavior from these models. Both these factors — the limited sample size and the discrepancy between the design of the algorithms and the characteristics of the dataset — could explain the observed results during the experiments on the MIMIC-IV dataset.

Moreover, both TARNet and CFR were outperformed by the S- and T-learners in terms of mean AUC. Table 2 shows that the S-learner achieved the highest mean AUC, followed closely by the T-learner. The high standard deviations for both learners, similar to those of TARNet and CFR, indicate that their performance also varies widely with different data splits or weight initializations. However, the maximum AUC values of the S- and T-learners are almost twice as high as those of TARNet and CFR. This suggests that they can achieve excellent performance under certain conditions. These results indicate that the metalearners would be more suitable for predicting the more appropriate PEEP regime for a patient's survival outcome.

5.3 Results from RCT experiments

The results from the RCT dataset experiments present a more nuanced picture, as seen in Table 3. The T-learner exhibited the highest AUC, suggesting that when applied to RCT data, its separate modeling of treatment and control groups may offer an advantage. This could be due to the randomized nature of treatment assignment in RCTs, which aligns well with the T-learner's assumption of independent treatment groups. The S-learner, while not achieving the highest AUC, still performed reasonably well, reinforcing its robustness across different types of datasets. Its performance, however, was not as dominant as in the MIMIC-IV experiments, indicating that the effectiveness of the S-learner may depend on the dataset characteristics and the nature of the treatment effects.

The performance of TARNet, with the second-highest AUC, suggests that it can capture complex relationships in RCT data (Table 3). The extremely low AUC of CFR was unexpected, as it had previously obtained a score better or equal to TARNet on both the simulations and the MIMIC-IV dataset. This result suggests that the balance regularization term may not translate as effectively to RCT data, where treatment assignment is already randomized and balanced. The wide confidence intervals (CI) for TARNet and CFR indicate substantial variance in their performance, which could be attributed to the limited sample size available for training the models. It is also worth noting that there is a significant overlap between the CIs of all four models, which may indicate the possibility that the differences in performance among the models may not be statistically significant.

5.4 Limitations

Reflecting on the research methodology, several limitations may have resulted from the decisions that could have influenced the outcome of the study. Firstly, the choice of imputation method, such as kNN, which assumes that similar patients have similar data, could have affected the distribution of the imputed values and, thus, the performance of the models. Different imputation strategies, like multiple imputation or model-based approaches, might have led to different conclusions about the effectiveness of the algorithms.

Secondly, the list of possible confounders from the features of the MIMIC-IV dataset was constructed based on previous studies [7,8,23] and empirical evidence. However, a more extensive analysis could reveal that another set of covariates ensures better results than the ones obtained during this project.

Thirdly, the choice of metrics and evaluation methods, especially the ones used for the MIMIC-IV and RCT datasets, may not fully capture the complexity of decision-making in ICU settings, where outcomes are multifaceted, and the cost of different types of errors can vary substantially. Alternative metrics or a combination of several could provide a more comprehensive evaluation.

Finally, the computational limitations encountered throughout the project resulted in an inability to perform extensive hyperparameter tuning for each scenario. This could have led to suboptimal configurations being used, especially for the more complex TARNet and CFR models, potentially affecting their performance. Future research could explore alternative methodologies to validate the results and potentially uncover new insights into the application of machine learning in healthcare settings.

6 Responsible Research

Analyzing ethical implications is crucial in research that intersects with healthcare, where the outcomes can directly affect patient well-being. This study used machine learning models to estimate ITE for mechanical ventilation in ICU patients. While these models can potentially improve patient care by personalizing treatment strategies, they also raise ethical concerns regarding fairness in treatment access. The models are trained on historical data, which may reflect existing biases or disparities in healthcare practices. If unaddressed, these biases could be perpetuated by the models, leading to unfairness in treatment recommendations. Therefore, it is an ethical obligation to scrutinize the data for such biases and to employ strategies that mitigate their impact before utilizing the methods in real-world scenarios, ensuring that the models serve all patient groups equitably.

Furthermore, the reproducibility of research methods and results is an important component of scientific integrity. In this project, steps have been taken to ensure that the methods employed are transparent and that the results can be independently verified. The experiments conducted are straightforward to recreate, with the data preprocessing, model training, and evaluation procedures thoroughly documented. Additionally, the source code has been made available in a public repository [31], and the datasets used are publically accessible [17], enabling other researchers to replicate this study.

7 Conclusions and Future Work

This project investigated the efficacy of neural network-based machine learning methods, specifically TARNet and CFR, in personalizing treatment strategies within intensive care units. The central research question of this study revolved around the potential of these methods to determine the individualized treatment effects of lower versus higher PEEP regimes on patient survival outcomes. The investigation was structured into three experimental stages: controlled simulation experiments, application to the real-world MIMIC-IV dataset, and validation through a randomized control trial dataset.

Simulation experiments were conducted to answer the first sub-question of this paper. They served as a foundational step, providing a controlled setting to rigorously test the models under diverse scenarios, assert their strengths and weaknesses, and construct expectations about their potential performance in real-world settings. The S-learner stood out for its consistently superior performance across all simulations. On the other hand, the results of the T-learner were less impressive, likely due to its separate modeling of treatment and control groups, which can be problematic in datasets with limited samples or unbalanced treatment. TARNet and CFR had similar results and often scored between the S-learner and the T-learner, benefiting from shared representation learning combined with treatment-specific modeling. However, they require a substantial amount of data to perform optimally.

The second sub-question was approached by analyzing the results of the estimators when applied to the MIMIC-IV dataset. It was noted the S-learner achieved the highest mean AUC score, as expected, given the results observed during the simulations. The T-learner also demonstrated robust performance. In contrast, the performance of TARNet and CFR was significantly lower. Moreover, TARNet's mean AUC was notably lower than that of CFR, implying that the balance regularization term incorporated in CFR may offer a significant advantage in unbalanced real-world datasets. CFR's performance was more consistent than TARNet's, although both displayed considerable variability, likely due to the limited sample size in the MIMIC-IV dataset. The RCT dataset provided an external benchmark, allowing for a comparative evaluation of the performance of the models. The randomness of the results, particularly for TARNet and CFR, highlighted the critical role of selecting the best model initialization to achieve reliable performance.

Recommendations for future research include evaluating the estimators on datasets with an extended number of samples, which would be crucial in further validating the effectiveness of the machine learning models. Moreover, exploring different imputation methods and confounder selection techniques could provide deeper insights into the models' performance and applicability in various medical settings. Additionally, employing a more comprehensive set of evaluation metrics could offer a more nuanced understanding of the models' decision-making processes and their implications in healthcare. Finally, it would be interesting to research how the performance of the models changes when evaluated on real-world datasets with continuous outcomes instead of binary survival outcomes.

References

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

[2] S. K. Sahetya and R. G. Brower, "Lung recruitment and titrated PEEP in moderate to severe ARDS: is the door closing on the open lung?" *JAMA*, vol. 318, no. 14, pp. 1327–1329, 2017.

[3] A. J. Walkey, L. D. Sorbo, C. Hodgson, N. K. J. Adhikari, H. Wunsch, M. O. Meade, E. Uleryk, D. Hess, D. 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.

[4] 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, P. Austin, S. Lapinsky, A. Baxter, J. Russell, Y. Skrobik, J. J. Ronco, and T. E. Stewart, "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*, vol. 299, no. 6, pp. 637–645, 02 2008. [Online]. Available: https://doi.org/10.1001/jama.299.6.637

[5] R. G. Brower, P. N. Lanken, N. R. MacIntyre, M. A. Matthay, A. H. 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, pp. 327–336, 2004.

[6] A. Mercat, J.-C. M. Richard, B. Vielle, S. Jaber, D. Osman, J.-L. Diehl, J.-Y. Lefrant, G. Prat, J. Richecoeur, A. Nieszkowska, C. Gervais, J. Baudot, L. Bouadma, L. Brochard, and f. t. Expiratory Pressure (Express) Study Group, "Positive end-expiratory pressure setting in adults with acute lung injury and acute respiratory distress syndrome: A randomized controlled trial," *JAMA*, vol. 299, no. 6, pp. 646–655, 02 2008. [Online]. Available: https://doi.org/10.1001/jama.299.6.646

[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, T. E. Stewart, and G. 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*, vol. 303, no. 9, pp. 865–873, 03 2010. [Online]. Available: https://doi.org/10.1001/jama.2010.218

[8] 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.

[9] J. Hoogland, J. IntHout, M. Belias, M. M. Rovers, R. D. Riley, F. E. Harrell Jr, K. G. M. Moons, T. P. A. Debray, and J. B. Reitsma, "A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint," *Statistics in Medicine*, vol. 40, no. 26, p. 5961–5981, Aug 2021.

[10] M. A. Hernán and J. M. Robins, *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC, 2020.

[11] 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*, vol. 116, no. 10, pp. 4156–4165, Feb. 2019.

[12] E. H. Kennedy, "Towards optimal doubly robust estimation of heterogeneous causal effects," *Electronic Journal of Statistics*, vol. 17, no. 2, Jan. 2023.

[13] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, Aug 2017, pp. 3076–3085.

[14] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, Jun. 2018.

[15] A. M. Alaa and M. van der Schaar, "Bayesian inference of individualized treatment effects using multitask gaussian processes," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[16] X. Nie and S. Wager, "Quasi-oracle estimation of heterogeneous treatment effects," *Biometrika*, vol. 108, no. 2, pp. 299–319, Sep. 2020.

[17] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[18] M. F. Alves, "Causal inference for the brave and true." [Online]. Available: https://matheusfacure.github.io/python-causality-handbook/landing-page.html

[19] A. Abadie and G. W. Imbens, "Matching on the estimated propensity score," *Econometrica*, vol. 84, no. 2, pp. 781–807, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA11293

[20] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*. PMLR, 2016, pp. 3020–3029.

[21] Y. Xie, J. E. Brand, and B. Jann, "Estimating heterogeneous treatment effects with observational data," *Sociological Methodology*, vol. 42, no. 1, pp. 314–347, 2012. [Online]. Available: https://doi.org/10.1177/0081175012452652

[22] "scikit-learn." [Online]. Available: https://scikit-learn.org/stable/index.html

[23] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets," *Journal of Biomedical Informatics*, vol. 83, p. 112–134, Jul. 2018.

[24] A. Curth and M. van der Schaar, "On inductive biases for heterogeneous treatment effect estimation," *Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[25] A. Curth, D. Svensson, J. Weatherall, and M. van der Schaar, "Really doing great at estimating CATE? A critical look at ML benchmarking practices in treatment effect estimation," *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[26] F. Devriendt, J. Van Belle, T. Guns, and W. Verbeke, "Learning to rank for uplift modeling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 10, pp. 4888–4904, 2020.

[27] "Tensorflow." [Online]. Available: https://www.tensorflow.org/

[28] A. Curth, "Catenets," Mar 2023. [Online]. Available: https://pypi.org/project/catenets/

[29] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," *arXiv preprint arXiv:1511.07289*, 2015.

[30] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, vol. 15, 2017.

[31] PetruAnica, "Cse3000-research-project." [Online]. Available: https://github.com/petruanica/cse3000-research-project

A Simulation details

Table 4: Details of the setups of the simulation experiments; $\beta_{2,4}$ in simulation 6 denotes beta distribution with parameters 2 and 4;

Sim. no.	*d*	*e(X)*	$\mu_0(X)$	$\mu_1(X)$	Remarks
1	10	0.1	$X^T\beta + 5\mathbb{I}(X_1 > 0.5)$	$\mu_0(X) + 8\mathbb{I}(X_2 > 0.1)$	$\beta \sim U([-5; 5]^d)$
2	10	0.5	$X^T\beta_1$	$X \cdot \beta_2$	$\beta_1, \beta_2 \sim U([1; 30]^d)$
3	10	0.5	$\frac{1}{2}\varsigma(X_1)\varsigma(X_2)$	$-\frac{1}{2}\varsigma(X_1)\varsigma(X_2)$	$\varsigma(x) = \frac{2}{1+e^{-12(x-0.5)}}$
4	10	0.5	$X^T\beta$	$\mu_0(X)$	$\beta \sim U([1; 30]^d)$
5	10	0.5	$\begin{cases} X^T\beta_{1-2} & \text{if } x_{10} < -0.4 \\ X^T\beta_{3-6} & \text{if } \lvert x_{10}\rvert \leq 0.4 \\ X^T\beta_{7-9} & \text{if } x_{10} > 0.4 \end{cases}$	$\mu_0(X)$	$\beta_{k-l} = \begin{cases} \beta(i) & \text{if } k \leq i \leq l \\ 0 & \text{otherwise} \end{cases}$ $\beta \sim U([-15; 15]^d)$
6	10	$\frac{1}{4}(1 + \beta_{2,4}(X_1))$	$2X_1 - 1$	$\mu_0(X)$	-