Individualized treatment effect prediction for Mechanical Ventilation

Using Causal Multi-task Gaussian Process to estimate the individualized treatment effect of a low vs high PEEP regime on ICU patients

Kieran McAlpine

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

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

Abstract

This research investigates the use of Causal Multi-task Gaussian Process (CMGP) for estimating the individualized treatment effect (ITE) of low versus high Positive End-Expiratory Pressure (PEEP) regimes on ICU patients requiring mechanical ventilation. The study addresses the complexities of determining ITE due to the inability to observe counterfactual outcomes and the confounding bias in observational studies. By employing Conditional Average Treatment Effect (CATE) estimators, such as S-Learner, T-Learner, and CMGP, the research evaluates the impact of different PEEP settings on patient survival across varied patient characteristics. The precision of these estimators is assessed using simulated data, real-world observational data from the MIMIC-IV dataset, and an external RCT dataset. The findings of this study are inconclusive, highlighting the need for further research to refine these methods and explore larger, more balanced datasets.

1 Introduction

Mechanical ventilation is crucial for critically ill patients with acute respiratory failure in the intensive care unit (ICU) [10]. About a third of the beds in the ICU in US hospitals are filled with patients that require mechanical ventilation [12]. Mechanical ventilation helps the patient by keeping the lungs properly aerated and keeping the oxygenation levels in the blood in order. One key setting of mechanical ventilation is the *Positive End-Expiratory Pressure* (PEEP) [1]. A higher PEEP regime will keep the alveoli in the lungs properly aerated, but could also cause unwanted extra damage to the lungs. Despite numerous trials, the debate between higher and lower PEEP remains [11]. Rather than a universal approach, it is believed that the benefits of higher PEEP might depend on patient characteristics.

This report has investigated the benefits of a higher vs lower PEEP regime using the *Individual Treatment Effect* (ITE). The ITE is a measure of how a certain individual within a population responds to receiving treatment compared to having not received treatment. Difficulties arise because an individual can only receive or not receive the treatment once. This means that only the factual outcome can be observed and not the counterfactual. Moreover, in observational/non-random studies, the treatment effect and selection into treatment can be intertwined. This may lead to confounding bias. Confounding bias can make correlational relationships between features seem like causative associations, which can lead to invalid conclusions in observational experiments. Figure 1 gives a generic example of confounding bias.

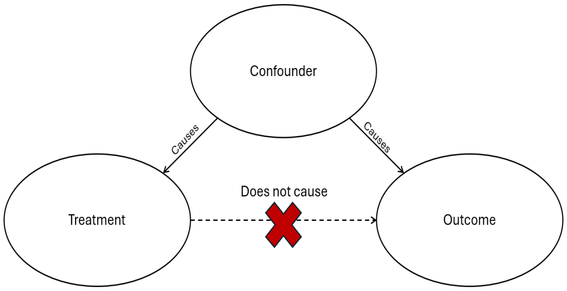

Figure 1: Own work. Example of a confounding feature in a diagram.

Normally this is where *Randomised Controlled Trials* (RCTs) come into play. By randomising treatments to certain patients, marginally on confounders, a good approximation of the counterfactuals can be made. In an ideal world it is also preferred to perform RCTs and observational tests side by side [5]. However the world is not ideal and performing RCTs is not always feasible, especially in the medical field. Patients might, understandably, not want to receive a random treatment assignment but the "best". This can cause sample sizes in RCTs to be too small to conclude anything meaningful.

Instead this report used the *Conditional Average Treatment Effect* (CATE) to infer the ITE [7]. CATE is similar to the ITE with the difference that instead of looking at an individual, the *Average Treatment Effect* (ATE) of a set of the population is examined. To this end CATE estimators are used. These are machine learning methods used in causal inference to estimate the ATE conditional on observed covariates. In simpler terms, it helps with understanding how the treatment (in this case, the choice between low and high PEEP) affects outcomes (such as survival) for different groups of patients. By estimating CATE, it can be determined if certain patients benefit more from one treatment over another, depending on their individual characteristics.

This research considered several CATE estimators, including S-Learner, T-Learner [9] and Causal Multi-task Gaussian Process (CMGP) [3]. The S-Learner and T-Learner both belong to the family of meta learners. These are models that are designed to work on top of base learners, which are existing machine learning algorithms that model the relationships between covariates and outcomes. CMGP works by treating the estimation as a multi-task learning problem, leveraging Bayesian nonparametric methods to handle observational data. Each of these CATE estimators provides a distinct approach and each has its strengths and weaknesses, for example CMGP starts to slow down for larger datasets because of an $O(n^3)$ matrix inversion. The precision of these CATE estimators was evaluated on simulated data, real world observational data and an external RCT dataset. The observational dataset is based on the MIMIC-IV dataset [8].

These CATE estimators address potential issues arising from the lack of complete and controlled randomization by making a set of accepted assumptions for the dataset. The first assumption is *unconfoundedness*, which ensures that there are no unobserved confounding factors influencing the selection into treatment. The second assumption is *common support*, which states that every individual, identified by a given set of covariates, has a non-zero probability of being observed in each of the treatment groups. Finally, the *Stable Unit Treatment Value Assumption* (SUTVA) asserts that the response to treatment of one individual is not affected by other individuals' treatment assignments and that the observed outcome is equal to the factual outcome, thereby ensuring no interference.

Within the context of using CATE estimation in medical research, this research seeks to answer the following question: *"How can Causal Multi-task Gaussian Process be used to estimate the individualized treatment effect of a low vs high PEEP regime on ICU patients?"* Accompanying this main research question are the following sub-questions:

1. How do S-Learner, T-Learner and Causal Multi-Task Gaussian Processes perform in estimating the individualized treatment effect of low vs high PEEP regimes?
2. What are the trade-offs between model complexity and computational efficiency among these methods?
3. Do the SUTVA, common support and unconfoundedness assumptions hold for the MIMIC-IV dataset?

The main related work is [3] which provided the knowledge and implementation for using Gaussian Processes in CATE estimation with the novel Causal Multitask Gaussian Process model. This also gave the inspiration for implementing the S-Learner and T-Learner with Gaussian Process Regression as a base learner. The findings of this research will contribute to gaining a better understanding of the effects on mortality of a high vs low PEEP regime on patients on mechanical ventilation in the ICU.

The paper is structured as follows. Section 2 describes the problem at hand. Section 3 will give more insight into the the methodology that was applied. Section 4 gives the experimental setup and results. Section 5 places the results in the broader context of the research field. Section 6 touches upon the ethical aspects of the research. This is followed by the conclusions in section 7. Finally section 8 discusses possible future research.

2 Problem description

This section gives some formal definitions and more concretely describes the difficulties with estimating the CATE on the observational ICU patient dataset specifically. This problem holds for multiple other real world use cases, although the medical field gets the most attention because of the high stakes associated with the decision to perform a treatment or not.

2.1 MIMIC-IV dataset

This research used the MIMIC-IV dataset [8]. This is a real world observational dataset of patients admitted to ICU in the United States between 2008 - 2019. This dataset houses the data of approximately 2900 patients, each with 24 features (age, sex, heart rate etc.). There is a treatment variable indicating if a high or low PEEP regime was chosen and finally there is an outcome variable which indicates the mortality of a patient 28 days after having received the treatment.

MIMIC-IV is an observational dataset. This means that the data has only been obtained after the fact and it is not known why an ICU doctor made a choice for a certain treatment. Also the dataset is unbalanced being skewed to substantially fewer patients having received a high PEEP regime treatment ($\approx 12\%$). Lastly the dataset also has confounding variables. These are variables that affect the decision to treat and the outcome variable. For example, whilst a higher age might influence a doctor to choose a certain PEEP regime, someone who is older is also indirectly more likely to pass away.

Some assumptions have to be made on the MIMIC-IV dataset before CATE estimation can be performed. These are unconfoundedness, common support and SUTVA.

The first assumption, unconfoundedness, makes sure that all confounders causing selection for treatment are observed in the data. This assumption is difficult to prove since it is only possible to reason about potential confounding factors and other potential confounders might have been missed if they were not recorded in the dataset to start with. In the MIMIC-IV dataset it can be reasonably argued that the most important confounders have been recorded, as this was the same patient data the ICU doctors had in front of them when choosing the PEEP regime for the patient. Looking at medical literature from the same period as when the data was recorded helps identify if recommendations for a treatment were given when certain patient features met certain values. Lastly there are other relevant covariates in the data that can act as a proxy for any possible missing confounders.

The next assumption, common support, states that each patient, with a set of covariates, has a non-zero probability of being observed with a high or low PEEP regime. Unlike the first assumption, common support can be inspected in the MIMIC-IV dataset. This was achieved by plotting density plots for each covariate and visually identifying if there are overlapping distributions between the the treated and untreated group. This can be observed in Figure 13 in Appendix D

The last assumption made, SUTVA, ensures that there is no interference between patients and that the observed outcome is equal to the factual outcome. In other words how a patient reacts to a high PEEP regime should not be affected by another patient's assignment to the high PEEP regime. Again this is difficult to check but it can be reasonably assumed that the treatment was given consistently across patients in the dataset.

2.2 Formal descriptions

This research is interested in estimating the ITE for an individual i. The two potential outcomes for i can be modeled as Y_i^1 when treatment has occurred and Y_i^0 if it has not. In this research receiving the high PEEP regime is equivalent to receiving treatment. From this it then follows that for i:

$$ITE = Y_i^1 - Y_i^0 \tag{1}$$

However as discussed in the introduction, one of the fundamental problems in causal inference is that it is impossible to know the counterfactual outcome of an individual. Only one of the two possible treatments $(Y_i^1, Y_i^0) \in \mathbb{R}^2$ is observed. Therefore the CATE was used. Firstly a patient i was defined in the dataset as $D_i = \{\mathbf{X}_i, Z_i, Y_i\}$. $\mathbf{X}_i$ is a set of covariates which are possible confounders that need to be controlled for, Z_i is the possible treatment assignment and Y_i is the possible outcome. For the MIMIC-IV dataset $Z_i, Y_i \in \{0, 1\}$ holds. The CATE is then defined as:

$$\tau(\mathbf{x}_i) = \mathbb{E}[Y_i^1 - Y_i^0 | \mathbf{X}_i = \mathbf{x}_i] \tag{2}$$

3 Methodology

In this section the mechanisms and theory behind the different CATE estimators - analysed in this research - are explained. As well as their advantages and disadvantages based on previous literature. Finally the experimental approach each CATE estimator was put through will be discussed.

3.1 Mechanism, advantages and disadvantages of the CATE estimators

Five CATE estimators were explored in this research. Four of these estimators are variations of the S-Learner and T-Learner frameworks, as described by [9], using different base learners: LGBMRegressor and Gaussian Process regression. The last estimator is the Causal Multitask Gaussian Process (CMGP), as described by [3].

S-Learner

The S-Learner combines the treatment and control data into a single model by augmenting the feature set with an indicator variable representing the treatment assignment. The model then predicts outcomes based on this augmented feature set. The estimated CATE can be represented as:

$$\hat{\tau}(\mathbf{x}_i) = \hat{f}(\mathbf{x}_i, 1) - \hat{f}(\mathbf{x}_i, 0) \tag{3}$$

Because an S-Learner only fits one regression / learner to the dataset, applying the same one to both the treated and untreated groups, it can get poor performance if there are big differences in the level of sparsity and smoothness between the treatment groups [2; 6]. For example if there is a big difference in outcome surface complexity between the groups, the S-Learner will perform poorly. However if the CATE is not complex the S-Learner will perform well. The S-Learner is also "easy" to implement and reason about and thus often used as a baseline.

T-Learner

The T-Learner builds separate models for the treatment and control groups. The difference between the predictions of these two models represents the CATE. The estimated CATE can be represented as:

$$\hat{\tau}(\mathbf{x}_i) = \hat{f}_1(\mathbf{x}_i) - \hat{f}_0(\mathbf{x}_i) \tag{4}$$

Unlike the S-Learner, the T-Learner fits two different regressions for the treated and untreated groups. This alleviates the problem of performing poorly when there is a big difference in outcome surface complexity. The T-Learner is also expected to do very well when the size of the data goes to infinity [2]. However, because the T-Learner splits the dataset into two, each individual regression has less data to train on, which can cause problems in accuracy if the dataset is very unbalanced and/or small. The T-learner can also not share any underlying information between the two groups as it estimates them independently, which can be detrimental in randomized studies where patients in the two groups can share the same distributional characteristics.

Causal Multi-task Gaussian Process

The Causal Multi-task Gaussian Process (CMGP), extends Gaussian Processes to handle multiple tasks simultaneously. It leverages the correlation between tasks (e.g., treatment and control) to improve the estimation of treatment effects by sharing information across related tasks. The estimated CATE can be represented as:

$$\begin{gathered}\hat{\tau}(\mathbf{x}_i) = \hat{f}_1(\mathbf{x}_i) - \hat{f}_0(\mathbf{x}_i) = \hat{\mathbf{f}}^\top(\mathbf{x}_i)\xi, \\ \text{where} \quad \xi^\top = [-1 \quad 1]\end{gathered} \tag{5}$$

CMGP is similar to the T-Learner, as it splits the data into the two subgroups for estimation. However due to the multitask approach of CMGP these two subgroups are not estimated independently. Instead CMGP is able to "combine" the optimization for estimation through some of the hyperparameters. This gives the advantages of both the T-Learner and S-Learner. An added bonus of CMGP is that it also gives individualized measures of confidence in the estimates through the posterior variance, making it more suitable for real life medical decisions.

Base learners

Both the S-Learner and the T-Learner are meta learners that use base learners. These base learners are machine learning models that already exist for other applications but have been augmented in their application by the meta learners to answer causal inference tasks. This research looked at two base learners for the S- and T-Learners. These are Gaussian Process Regression (GPR) and Light Gradient Boosting Machine Regression (LGBMR). GPR was chosen so that the difference between the single output regression models of the meta learners and the multi output regression model of CMGP can be analysed while keeping the type of regression similar. The downside of GPR is that it can run into computational issues if the datasets become too large due to a $O(n^3)$ matrix inversion, where n is the size of the covariance matrix. LGBMR was chosen as it is widely used and does not have as many performance issues as GPR.

Kernels and hyperparameters

Kernels need to be defined for the Gaussian Processes. A kernel (or covariance function) defines the covariance between any two points in the input space. A *Radial Basis Function* (RBF), also known as a *Squared Exponential Function* was used for the GPR. The function for this kernel is defined as:

$$\begin{gathered}K_{\text{RBF}}(x, x`) = \exp\left(-\frac{|d|^2}{2l^2}\right), \\ \text{where} \quad d = x - x` \\ \text{and } l \text{ is the characteristic length-scale}\end{gathered} \tag{6}$$

RBF is a stationary kernel that is infinitely differentiable, making it very smooth. The length-scale parameter controls how quickly the correlation between points decrease with distance. The RBF kernel was chosen as it is commonly used in Gaussian Processes and is also used when making the kernel for the *Linear Coregionalization Model* (LCM) in CMGP. To allow the model to capture the variations in the data more accurately, a separate length-scale was set for each dimension. This approach is known as *Automatic Relevance Determination* (ARD). CMGP also has a subroutine built in that initializes more hyperparameters, such as the signal and noise variance. For simplicity this was not done for the GPR base learners.

The CATE estimators utilizing Gaussian Process are optimized with a maximum of 100 iterations using a built-in *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) optimizer. The CATE estimators with LGBMR as a base learner are fitted with a maximum depth of 2 and 20 minimum child samples. All other (hyper)parameters are default values as defined by the GPy [1] and LightGBM [2] python libraries.

3.2 Experimental Approach

The selected CATE estimators have been applied in three separate settings. Firstly, they have been validated and compared using simulated datasets. Secondly, they have been trained on a part of the MIMIC-IV dataset and used to make predictions on the other part. Lastly, the CATE estimators have been trained on the entire MIMIC-IV dataset and used to perform CATE prediction on an external RCT dataset.

Simulation

Since there is no access to the counterfactual in real-world data, simulating data according to predefined functions is a popular way of validating and comparing models. The simulation can calculate the factual and counterfactual outcomes, thus providing the true CATE. The models are then trained while withholding the counterfactuals, predict the CATE for the simulated data, and then are validated by comparing the predicted CATE with the true CATE. To determine the accuracy of the models, the average $n = 10$ *Mean Squared Error* (MSE) between the predicted CATE and true CATE for increasing training set sizes [200, 500, 1000, 1500, 2000] has been plotted.

Three different simulations have been run based on parameters described in [9]. Each simulation provides a different scenario. The first simulation emulates an unbalance in treatment assignment, the second simulation emulates confounding features and lastly the third simulation emulates an unbalanced and confounded dataset similar to the MIMIC-IV dataset. The exact simulation parameters can be found in Appendix A.

Real-World Data

Before the models could be executed on the MIMIC-IV dataset, preprocessing steps have been undertaken. Firstly, two columns with metadata, *‘id’* and *‘Unnamed: 0’*, have been removed. Categorical features have then been encoded to numerical ones:

- *sex*: ‘F’ has been mapped to 0 and ’M’ to 1
- *mort_28*: ‘False’ has been mapped to 0 and ’True’ to 1
- *peep_regime*: ‘low’ has been mapped to 0 and ’high’ to 1

After remapping the data, the next step of preprocessing involved normalizing the data using the standard normal scaler, with all (hyper)parameters set to the default values as defined by the scikit-learn[3] python library. Gaussian Processes often require the individual features to look like standard normally distributed data. This is why, for example, Min-Max scaling was not chosen. Lastly, a KNN-Imputer was used to estimate missing values within the dataset. A KNN-Imputer imputes missing data by taking the mean of the m nearest neighbors found in the training dataset. For this imputation, $n = 5$ was chosen, with all other (hyper)parameters set to the default values as defined by the scikit-learn Python library. Imputation was chosen, as dropping the rows with empty values would have yielded a dataset about half the size of the original dataset. This would have been approximately 1500 rows, which would have been too small a sample for the models.

Next, the covariates in the dataset that are possible confounders were selected. This was achieved by exploring which covariates are good at predicting outcome and treatment and/or are mentioned in other research. This approach resulted in the following list of potential confounders: *’age’, ’pf_ratio’, ’po2’, ’pco2’, ’fio2’, ’hco3’, ’peep’, ’plateau_pressure’, ’respiratory_rate’, ’weight’, ’driving_pressure’*.

The dataset was split into 70% train and 30% test sets. The test set was then calibrated / debiased. The calibrating was done using a Random Forest, with 50 estimators, 20 minimum leaf samples and a max depth of 2, to estimate the propensity score. The debiasing was done as described by [4]. The calibrated data is depicted in Figure 12.

The performance of the models was analysed using a cumulative gain curve and the area under it. The cumulative gain curve is a graphical representation of how well the model is able to identify the positive responders for a treatment compared to a random selection. The x-axis represents, according to the model, from left to right the top 10% responders, the top 20 % and so on. The y-axis represents the cumulative percentage of positive instances identified by the model. “Positive” instances are true instances of mortality after 28 days, thus indicate death. The cumulative gain curve was implemented as by [4].

External validation

Aside from the MIMIC-IV dataset there was also a RCT dataset used for external validation. Therefore the models were trained on the complete MIMIC-IV dataset and then these saved models were sent off for external validation. Similarly to the previous section the performance of the models was analysed using cumulative gain curves and the area under it.

[1] https://gpy.readthedocs.io/en/devel/index.html

[2] https://lightgbm.readthedocs.io/en/latest/index.html

[3] https://scikit-learn.org/stable/index.html

4 Experimental setup and results

This section provides and interprets the results of the experiments and gives the environment variables which the experiments were ran on. All the experiments were run on a laptop with an Intel(R) Core(TM) i7-9750H CPU @ 2.60 GHz, 16 GB Memory @ 2667 MHz, NVIDIA Quadro P2000 and Windows 11 Home.

4.1 Simulation results

This section provides the results of the simulation experiment, that emulates the MIMIC-IV dataset. Figure 2 plots the average MSE of each CATE estimator against the training set size. Figure 3 plots the execution times of each CATE estimator against training set size. Table 1 provides the mean MSE and standard deviation of the CATE estimators for the largest training size. Results for the other simulations can be found in Appendix B.

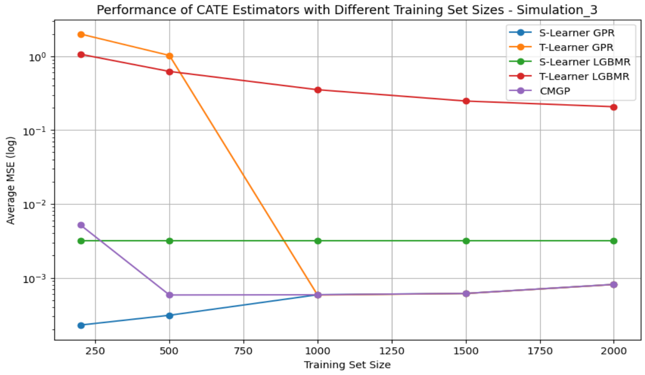

Figure 2: Average (n=10) MSE for each CATE estimator for increasing training set sizes. Lower is better. Simulation 3 emulates the MIMIC-IV dataset.

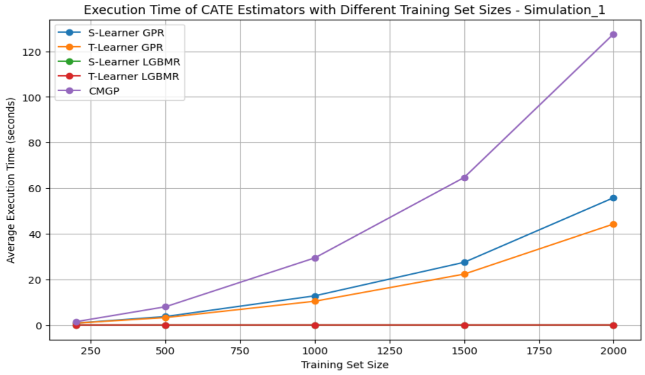

Figure 3: Average (n=10) execution time (seconds) for each CATE estimator for increasing training set sizes. Lower is better.

Model	Mean	Standard Deviation
S Learner GPR	8.13e-04	1.44e-04
T Learner GPR	8.11e-04	1.46e-04
S Learner LGBMR	3.21e-03	0.0
T Learner LGBMR	2.07e-01	5.13e-02
CMGP	8.13e-04	1.44e-04

Table 1: Mean and Standard Deviation of the MSE for Simulation 3 for different models with train size 2000. Lower is better.

4.2 MIMIC-IV results

This section presents the results of running the CATE estimators on the MIMIC-IV dataset. Cumulative gain curves for each different CATE estimator with their estimations for the test and train sets have been provided. A large difference between these two lines indicates that the estimator is overfitting on the training data. Additionally, a table with the normalized area between random allocation and the predictions for the test set has been provided, where a larger score is better.

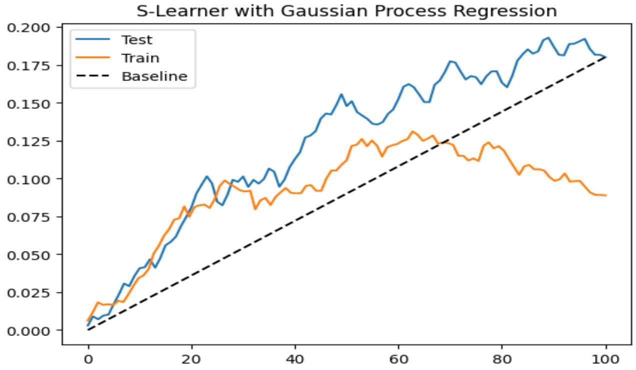

Figure 4: Cumulative gain curve for S-Learner with GPR. x-axis: Cumulative Population (%). y-axis: Cumulative Gain (%).

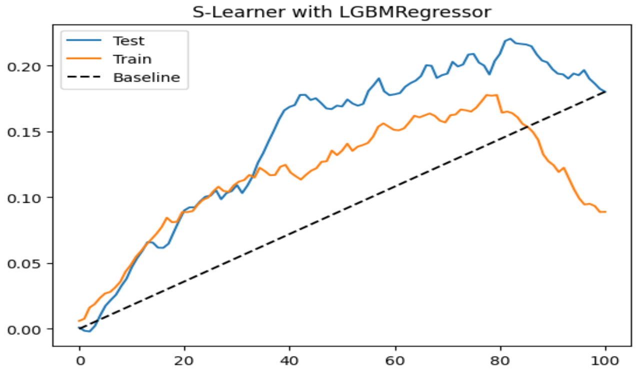

Figure 5: Cumulative gain curve for S-Learner with LGBMR. x-axis: Cumulative Population (%). y-axis: Cumulative Gain (%).

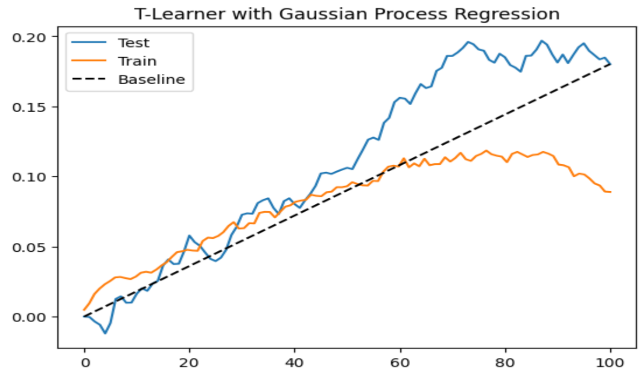

Figure 6: Cumulative gain curve for T-Learner with GPR. x-axis: Cumulative Population (%). y-axis:Cumulative Gain (%).

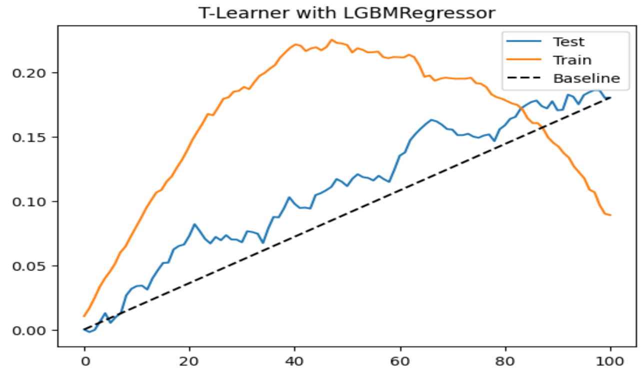

Figure 7: Cumulative gain curve for T-Learner with LGBMR. x-axis: Cumulative Population (%). y-axis: Cumulative Gain (%).

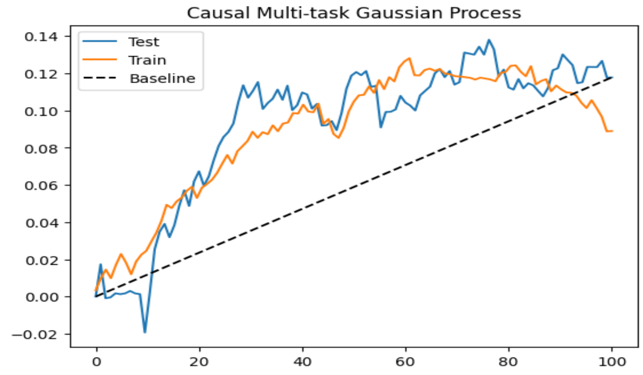

Figure 8: Cumulative gain curve for CMGP. x-axis: Cumulative Population (%). y-axis: Cumulative Gain (%).

Model	Normalized ROC-AUC score
S-Learner with GPR	3.415e-02
T-Learner with GPR	2.151e-02
S-Learner with LGBMR	5.601e-02
T-Learner with LGBMR	1.967e-02
CMGP	3.314e-02

Table 2: ROC-AUC scores for the cumulative gain curves for the different CATE estimators. Larger is better.

4.3 RCT results

This section provides the results of the models performing CATE estimation on the external validation dataset after having been trained on MIMIC-IV.

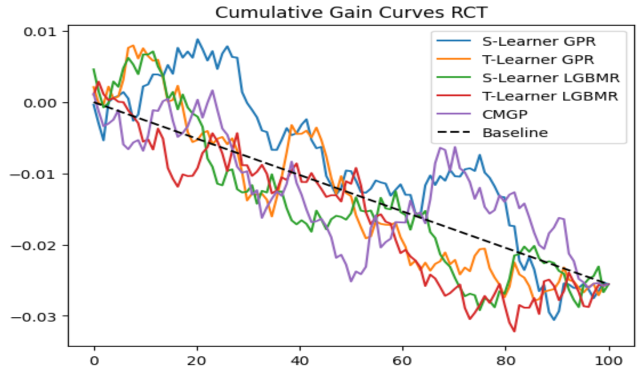

Figure 9: Cumulative gain curves for all models. x-axis: Cumulative Population (%). y-axis: Cumulative Gain (%)

5 Discussion

This section analyses the results from the experiments provided in section 4 and looks critically at the methodology described in section 3.

5.1 Simulation

In the simulations it can be observed that the S-Learner with GPR and CMGP perform the best from the start but do not improve significantly. The other CATE estimators perform poorly at the start - especially the T-learners - and improve when the training set size increases. This is the expected behaviour as it already has been mentioned that T-Learners perform badly in unbalanced and small datasets. Another interesting observation is that in nearly all simulations the meta learners with GPR as a base learner outperforms the meta learners with LGBMR.

As expected the CATE estimators with Gaussian Processes experience significant slow down when increasing the training set size. CMGP performed the worst because it has a more complex kernel to optimize with more dimensions. This is followed by the S-Learner with GPR and then the T-Learner with GPR, that performed slightly better because it splits the dataset into two separate optimization tasks and does not add an extra dimension (treatment variable). The meta learners with LGBMR performed significantly better. The matrix inversion slow down could dissuade usage of Gaussian

Processes for larger or high dimensional datasets. However for smaller datasets the run time is similar and mean MSE is smaller, making them attractive for this usecase. Also datasets - even observational - are usually smaller in the medical field, like MIMIC-IV. Moreover the execution time is not cripplingly slow. There are also other methods of speeding up the execution time that were not examined in the methodology. These are using Sparse Gaussian Process Regression methods or down sampling the training set, like is already done in example implementations of CMGP.

5.2 MIMIC-IV

From the cumulative gain curves, it is possible to check for overfitting in the models and compare them based on the area under the test gain curve. All meta learners show some degree of misfitting, with the T-Learner using LGBMR overfitting the most. Figures 4, 5, and 6 initially show the test and train gain curves in close proximity, indicating a good fit. However, the train cumulative gain curve stagnates or dips compared to the test gain curve when reaching the top 60% - 80% responders according to the model. Figure 7 shows misfitting throughout the entire dataset by the model.

The T-Learners show the most misfitting, which is expected as the MIMIC-IV dataset is unbalanced and not very large. This imbalance causes the T-Learners to fit or optimize on a relatively small treated group, potentially leading the models to learn noise and therefore overfit. As shown in Figure 8 CMGP does not exhibit overfitting, with the test and train gain curves closely matching each other. CMGP handles overfitting the best, incorporating *Leave-One-Out Cross-Validation* (LOO-CV) to evaluate the empirical error in factual outcomes. LOO-CV helps in assessing how well the model generalizes to new data points, providing a check against overfitting.

The cumulative gain curve slopes upwards indicating that for a higher PEEP regime there are more instances of mortality. Comparing the normalized AUC scores between the models in Table 2 shows barely any difference among them, with S-Learner LGBMR slightly ahead of the other models. The models that perform worse than the rest are the two T-Learners. These outcomes were also present in the simulations, i.e. most models having the same performance and the T-Learner with LGBMR performing slightly worse.

Nevertheless the other models do not appear to show significantly better performance than the baseline random allocation, with all showing a relatively a low ROC - AUC score.

5.3 RCT

The RCT results provide a different conclusion than what arose from predicting on the MIMIC-IV dataset. Instead of the cumulative gain curves sloping upwards, the curves and baseline slope downwards. This indicates that a high PEEP regime has less responders to mortality. In other words a high PEEP regime results in less deaths. However simlar to the MIMIC-IV results the ROC-AUC scores are very low and the models all hover around the random allocation baseline. Only S-Learner with GPR and CMGP appear to perform better than random.

6 Responsible Research

This section describes the steps taken to address concerns regarding the results collected, the reproducibility of the experiments conducted, and the MIMIC-IV patient data.

By demonstrating how these issues have been addressed, this research aims to foster trust in the findings and contribute to the body of knowledge in the field of ICU treatment research. This commitment not only enhances the credibility of the study but also enables other researchers to build upon this work, advancing the understanding of CATE estimation in critical care settings.

6.1 Data Collection

Rigorous data collection and analysis practices have been adhered to, ensuring the integrity and reliability of the findings. The data used in this study comprises both real-world anonymized patient data and simulated data, allowing for a comprehensive analysis while maintaining patient confidentiality.

To mitigate the impact of outliers and reduce variance in the results, a robust methodology has been adopted where the results of multiple predictions have been averaged. This approach helps to stabilize the estimates, providing a more accurate representation of the treatment effects across different PEEP regimes.

6.2 Transparency and Reproducibility

To facilitate reproducibility and ensuring that the findings can be independently verified, comprehensive documentation of all implementation details has been provided. This includes the specifics of data preprocessing, model training, and evaluation processes. The code and data processing scripts have been made available to the research community[4], allowing others to replicate the study and validate the results.

6.3 MIMIC-IV

Given the sensitive nature of real-world patient data, the MIMIC-IV database already implements stringent measures to address privacy concerns [8]. Firstly, access to the data is restricted to approved individuals after completing an online course covering important aspects of research with human participant data; *CITI Data or Specimens Only Research*[5]. Additionally, all patient data has already been rigorously anonymized to ensure that individual identities are protected. The anonymization process involves removing or encrypting all personally identifiable information (PII) to prevent any possibility of re-identification. The dataset has also been reviewed by an Institutional Review Board at the Beth Israel Deaconess Medical Center, which has granted a waiver of informed consent and approved the data sharing initiative [8].

Moreover, the models are designed solely to estimate treatment effects and do not attempt to infer or predict the identities of the patients. The focus remains strictly on understanding the impact of different PEEP regimes on patient outcomes, without compromising patient privacy. All relevant

[4] https://github.com/kierma/CSE3000-CATE-estimators

[5] https://about.citiprogram.org/

data protection regulations and ethical guidelines have been adhered to, safeguarding the confidentiality of the data.

7 Conclusions

This section provides the final conclusions on the main research question and its accompanying sub questions.

How can Causal Multi-task Gaussian Process be used to estimate the individualized treatment effect of a low vs high PEEP regime on ICU patients? It cannot be concluded whether a CMGP can be used to to predict a high or low PEEP regime, as the cumulative gain curves of the external dataset and test set contradict each other. In addition the normalized areas are too small to draw any definite conclusions.

How do S-Learner, T-Learner and Causal Multi-Task Gaussian Processes perform in estimating the individualized treatment effect of low vs high PEEP regimes? The S-Learner with GPR and CMGP performed the best in the simulations, MIMIC-IV predictions and RCT predictions.

What are the trade-offs between model complexity and computational efficiency among these methods? There is significant performance degradation in the CATE estimators using Gaussian Processes, which may dissuade usage for large datasets / datasets with a high dimension. The T-Learners performed the worst.

Do the SUTVA, common support and unconfoundedness assumptions hold for the MIMIC-IV dataset? The three assumptions - unconfoundedness, common support and SUTVA - can reasonably be assumed to hold for the MIMIC-IV dataset.

8 Future Work

There are some interesting aspects that can be looked at in further research. Firstly the Gaussian Process base learners have currently been implemented with quite simplistic RBF kernels. Kernel selection is important and can result in varying levels of performance. Also investigating the use of sparse GPR can be interesting if the excessive run time is a hindrance.

Secondly this research only investigated GPR as a base learner in the two most common meta learners, S and T. However there is a whole swathe of other meta learners such as X, R, DR etc. that might benefit from having GPR as a base learner. Especially since the simulations showed that in general the meta learners with GPR outperformed their LGBMR counterparts in terms of mean MSE.

Thirdly there is still quite a lot unknown about possible confounders for the PEEP regime. Only a handful of confounders are quite certain, with the rest being educated guesses based on domain knowledge. Further research could more robustly try to identify possible confounders.

Lastly the MIMIC-IV dataset was quite a small dataset - $\approx$ 2900 rows - making it difficult to reliably draw conclusions without just attributing a result to a model learning noise. A larger dataset could help with this and provide more accurate / better results.

A Simulation parameters

Simulation 1

$$\begin{aligned}
&e(x) = 0.1, \quad d = 20, \\
&\mu_0(x) = x^T\beta + 5\mathbb{I}(x_1 > 0.5), \quad with \quad \beta \sim Unif([-5, 5]^{20}), \\
&\mu_1(x) = \mu_0(x) + 8\mathbb{I}(x_2 > 0.1)
\end{aligned}$$

Simulation 2

$$\begin{aligned}
&e(x) = \frac{1}{4}\left(1 + \beta(x_1, 2, 4)\right), \quad d = 20, \\
&\mu_0(x) = 2x_1 - 1, \\
&\mu_1(x) = \mu_0(x),
\end{aligned}$$

Simulation 3

$$\begin{aligned}
&d = 24, \\
&e(x) = \frac{1}{1 + e^{-L^T\beta}}, \\
&e_{\text{unbalanced}} = \frac{0.12}{\text{mean}(e(x))} \cdot e(x), \\
&where \quad L = \{x_1, x_3, x_4, x_5, x_9, x_{15}, x_{16}, x_{20}, x_{21}, x_2, x_6\}, \\
&\mu_0(x) = \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_4 + \beta_4 x_5 + \beta_5 x_9 + \beta_6 x_{15} \\
&\qquad + \beta_7 x_{16} + \beta_8 x_{20} + \beta_9 x_{21} - 1, \\
&\mu_1(x) = \mu_0(x) + \beta_{10} x_{19}
\end{aligned}$$

B Simulation results

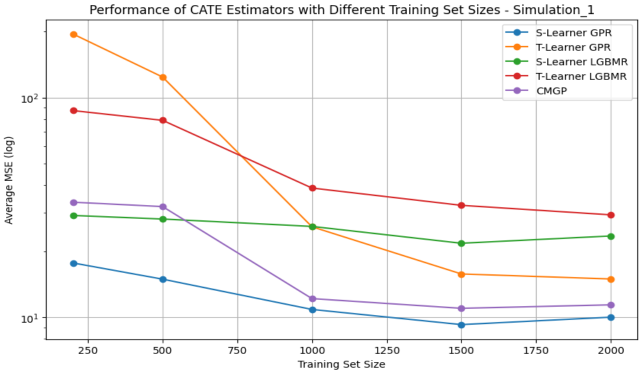

Figure 10: Average (n=10) MSE for each CATE estimator for increasing training set sizes. Lower is better. Simulation 1 emulates an unbalanced treatment assignmen.t

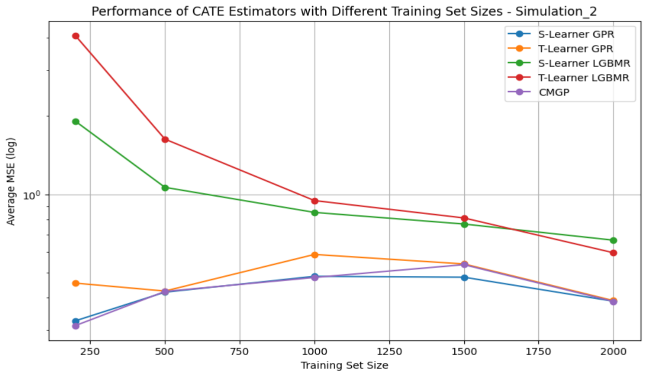

Figure 11: Average (n=10) MSE for each CATE estimator for increasing training set sizes. Lower is better. Simulation 2 emulates confounding features.

Model	MSE	Standard Deviation
S Learner GPR	1.00e+01	9.57e-01
T Learner GPR	1.49e+01	1.93e+00
S Learner LGBMR	2.35e+01	3.85e+00
T Learner LGBMR	2.93e+01	4.61e+00
CMGP	1.14e+01	4.70e+00

Table 3: Mean and Standard Deviation of the MSE for Simulation 1 for different models with train size 2000. Lower is better.

Model	Mean	Standard Deviation
S Learner GPR	3.88e-01	6.49e-02
T Learner GPR	3.90e-01	6.53e-02
S Learner LGBMR	6.66e-01	5.55e-02
T Learner LGBMR	5.96e-01	9.56e-02
CMGP	3.87e-01	6.35e-02

Table 4: MSE and Standard Deviation of the MSE for Simulation 2 for different models with train size 2000. Lower is better.

C Debiasing MIMIC-IV

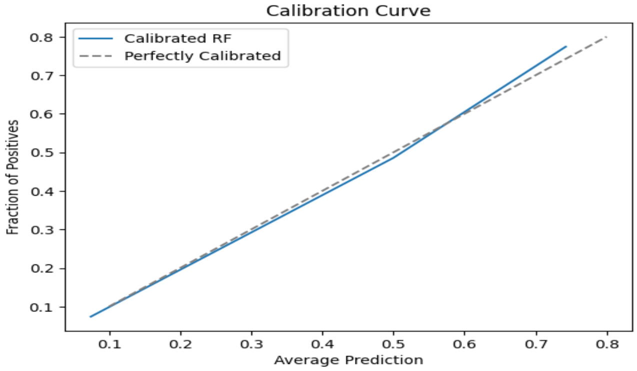

Figure 12: How well the debiased test set is calibrated. Closer to the dashed line is better.

D Density plots

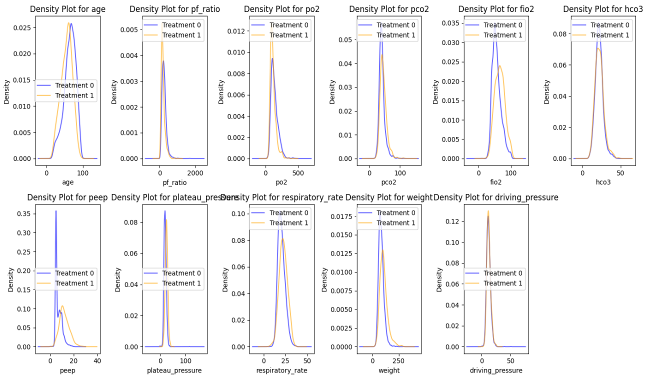

Figure 13: Density plots for the features that are possibly confounders. A bigger overlap is better.

References

[1] Pilar Acosta, Edgardo Santisbon, and Joseph Varon. The use of positive end-expiratory pressure in mechanical ventilation. *Critical Care Clinics*, 23(2):251–261, April 2007.

[2] Ahmed Alaa and Mihaela van der Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 129–138. PMLR, 10–15 Jul 2018.

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

[4] Matheus Facure and Michell Germano. matheusfacure/python-causality-handbook: First edition, 2021.

[5] David Faraoni and Simon Schäfer. Randomized controlled trials vs. observational studies: Why not just live together? *BMC Anesthesiology*, 16, 12 2016.

[6] P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, 15(3):965 – 2020, 2020.

[7] Daniel Jacob. Cate meets ml – the conditional average treatment effect and machine learning, 2021.

[8] Alistair Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei Lehman, Leo Celi, and Roger Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 01 2023.

[9] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, February 2019.

[10] Martin Tobin. Advances in mechanical ventilation. *The New England journal of medicine*, 344:1986–1996, 06 2001.

[11] 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.

[12] Hannah Wunsch, Jason Wagner, Maximilian Herlim, David H. Chong, Andrew A. Kramer, and Scott D. Halpern. Icu occupancy and mechanical ventilator use in the united states*. *Critical Care Medicine*, 41(12):2712–2719, December 2013.