MACHINE LEARNING AND CAUSAL INFERENCE FOR THE ESTIMATION OF THE EFFECT OF TACROLIMUS ON KIDNEY REJECTIONS

Inês Borges Carioca Moreno Rodrigues

Student id: 5552621

Pattern Recognition & Bioinformatics Research Group
Thesis Advisor: Prof. Marcel Reinders
Daily Supervisor: Jesse Krijthe

A thesis submitted for the degree of Masters of Computer Science - Artificial Intelligence Track

Delft
19th of July, 2023

Machine Learning and Causal Inference for the estimation of the effect of tacrolimus on kidney rejections

Inês B.C.M. Rodrigues[1]

[1]Pattern Recognition & Bioinformatics group, TU Delft, Delft

Abstract

Tacrolimus is an immunosuppressive drug given to kidney transplant patients. A low concentration of this drug can lead to kidney rejection, but to our knowledge no research has been done to causally connect the two. This paper investigates the causal effect of tacrolimus concentration on kidney rejection occurrence using predictive analysis and a marginal structural model. The data utilized in this study was obtained from a randomized clinical trial conducted at the Erasmus Medical Center, Rotterdam. The challenges posed by limited data availability and class imbalance were carefully considered in designing the model structures. To investigate the predictive properties of tacrolimus related variables we compared results of Logistic Regression and XGBoost models on different sets of variables, yielding inconclusive results. To measure the causal effect of tacrolimus concentrations on the rejection probability, a marginal structural model was developed to estimate the causal effect of the percentage of hours spent within the target tacrolimus concentration range on the probability of kidney rejection. While a large amount of uncertainty remains, our estimates tentatively indicate a decrease as the percentage in rejection probability as the percentage of hours on target increased. Future studies are recommended to explore alternative datasets to enhance the confidence of the findings.

1 Introduction

When receiving a kidney transplant, patients are administered immunosuppressive therapy to prevent rejection of the newly transplanted organ. One key component of this therapy is tacrolimus, an immunosuppressive agent known for its potent immunosuppressive properties [Hooks, 1994].

Tacrolimus is not without potential adverse effects, as high concentrations can lead to nephrotoxicity or increase the incidence of diabetes, while low concentrations may fail to adequately suppress the immune response, potentially resulting in kidney rejection. Hence, close monitoring and individualized adjustment of tacrolimus concentrations are crucial for optimal therapeutic outcomes. However, studies investigating the precise relationship between tacrolimus exposure and the risk of acute rejection [Israni et al., 2013], [Gatault et al., 2017], [Bouamar et al., 2013] have yielded inconsistent findings, often attributed to variations in therapeutic protocols or limited sample sizes[Brunet et al., 2019].

In collaboration with the Erasmus MC Transplant Institute at Erasmus University Medical Center, Rotterdam, the present study aims to investigate the impact of immunosuppressive treatment on kidney allograft rejection in kidney transplant recipients. Specifically, our objective is to elucidate the causal effect of tacrolimus concentration on the occurrence of kidney rejection, thereby contributing to a more comprehensive understanding of the effect of this drug on patient's bodies.

The link between immunosuppressive therapies and kidney rejection has been explored in previous investigations, employing diverse methodologies such as statistical studies[Israni et al., 2013] [Gatault et al., 2017] [Bouamar et al., 2013], pharmacokinetic modeling [Mould and Upton, 2013], and machine learning techniques [Truchot et al., 2022]. However, the application of observational causal inference methods to elucidate the treatment-outcome relationship in this context remains largely unexplored.

This study endeavors to advance our understanding of this treatment-outcome relationship and quantify its impact through a multifaceted approach. First, a causal graph was constructed to provide a comprehensive depiction of the intricate causal relationships at play. Subsequently, predictive analyses were conducted using classification models and feature selection techniques to establish the predictive value of specific variables for the desired outcome. Finally, causal inference analysis was performed through a marginal structural model to estimate the average treatment effect of tacrolimus concentration on the occurrence of kidney rejection in patients.

This integrative approach has the potential to enhance our understanding of the causal mechanisms underlying treatment response, ultimately facilitating improved therapeutic strategies in the field of kidney transplantation.

The paper is structured as follows: the first two sections will go through related and the data. We will then go through the methods employed for both the predictive and causal analysis. The implementation and results of the predictive analysis will then be presented, followed by the causal analysis. Finally, we will discuss these results in the Discussion.

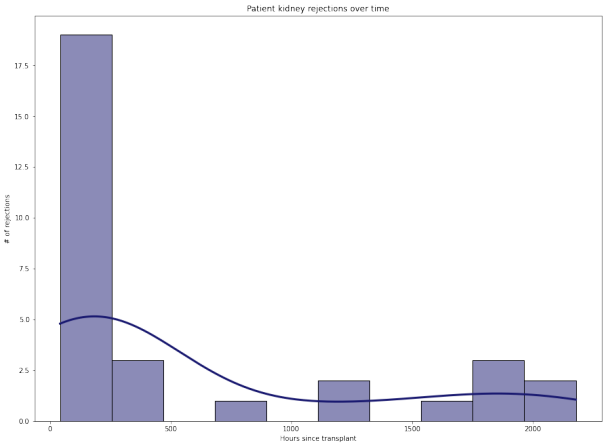


Figure 1: Number of rejection events over time. A majority of rejections happen before hour 250, meaning before 10 days have passed since the transplant.

2 Related Works

In [Sapir-Pichhadze et al., 2014] the authors use time-dependent Cox proportional hazards models to investigate the association between the standard deviation of tacrolimus levels starting at 1-year post-transplant and kidney rejection. The authors found a significant 27% increase in the adjusted hazard of rejection for every 1-unit increase in tacrolimus standard deviation.

Traditional statistical methods have been employed to predict risk of kidney rejection in [Loupy et al., 2019]. Here the authors used univariable Cox regression analyses and fractional polynomial methods to assess associations between allograft failure and various factors. A risk prediction score (iBox) was then calculated for each patient based on regression coefficients. This prediction score, along with other traditional statistical methods is pitted against Machine Learning techniques for the prediction of kidney rejection in [Truchot et al., 2022]. The authors included a cohort of over 8000 patients from diverse populations, employing machine learning survival models such as Random Survival Forests, XGBoost, Support Vector Machines, and others, alongside the well-established iBox and other Cox proportional hazards models. The study concluded that overall Machine Learning methods do not outperform traditional statistical methods. Although their analysis and our study differ in model objectives, data characteristics, and follow-up durations, it is useful to understand how rejection has been predicted in the past. The absence in our study of variables found influential in [Truchot et al., 2022] (ex: estimated glomerular filtration rate) warrant exploration in future studies with shorter follow-up periods such as ours. On the other hand, these two studies predict kidney rejection risk by only taking into account transplant compatibility and demographic factors, whereas we study the predictive association and causal effect of treatment for the same outcome.

Our causal analysis makes use of the Marginal Structural Model (MSM) method [Robins et al., 2000]. This approach has demonstrated its effectiveness in various studies. In [Hernán et al., 2000], the authors use a MSM to estimate the effect of zidovudine therapy on mean CD4 count among HIV-infected men. For this specific objective, a Cox proportional hazards model was deemed suitable. However, a key challenge arose due to the temporal variation of patient weights, which was not accommodated by most software packages that lacked support for patient-specific time-varying weights. To overcome this limitation, the authors redefined their survival model as a simplified weighted pooled logistic regression. This alternative model predicted the probability of death at each time point t for all surviving patients, while conditioning on prior treatment and covariates. The authors argued that the results obtained from this modified approach remained comparable to those derived from the Cox model, given the relatively small hazard associated with any individual month. Our method takes this approach as inspiration to adapt the classic MSM to our setting where weights vary over time and patient.

In [Shahn et al., 2020], data from the initial 24 hours of ICU admission was employed to predict mortality at a 30-day interval. In this scenario, all patients were consistently monitored for the same duration (24 hours) to anticipate a future event.

3 The Data

Data was obtained from a randomized clinical trial [Shuker et al., 2016] performed at the Erasmus Medical Center, Rotterdam, where adults received a single-organ, blood group ABO-compatible kidney from a living donor. Patients were randomly assigned to either receive tacrolimus in a standard, body-weight-based dose of 0.2 mg/kg/day or a dose dependent on their CYP3A5 genotype. During hospitalization tacrolimus was taken at 10:00 and 22:00 h and patients were instructed to continue doing so after discharge.

Two-hundred-forty patients were included and randomized. Out of these 231 patients completed the 3 months of follow-up. Clinical endpoints considered were incidence of biopsy-proven acute rejection(BPAR), the incidence of clinically presumed acute rejection, and renal function at month 3 after transplantation. In case of rejection the patients were censored at time of rejection.

In total, 40 variables were measured. Of these, 33 were taken either before the transplant or at the start of treatment and are static. Among the static variables, 11 were excluded because their effect was described in other variables. For example, mmA (HLA-A missmatches), mmB (HLA-B missmatches), and mmDR (HLA-DR missmatches) were excluded because of the use of mmTotal, the sum of mmA, mmB, and mmDR. Over the course of treatment, 7 other variables were measured: tacrolimus concentration, tacrolimus dosage, Mycophenolic acid (MPA) concentration, MPA dosage, creatinine, hematocrit (hct), and albumin. Because of their time-varying characteristic, these will be referenced as dynamic variables. Static variables with missing values were imputated with the mean of the values of the 5 nearest neighbours (5-NN imputation).

Because tacrolimus concentration is a continuous variable, there was a need to measure the levels of tacrolimus in a simpler way. This happens mainly in causal analysis

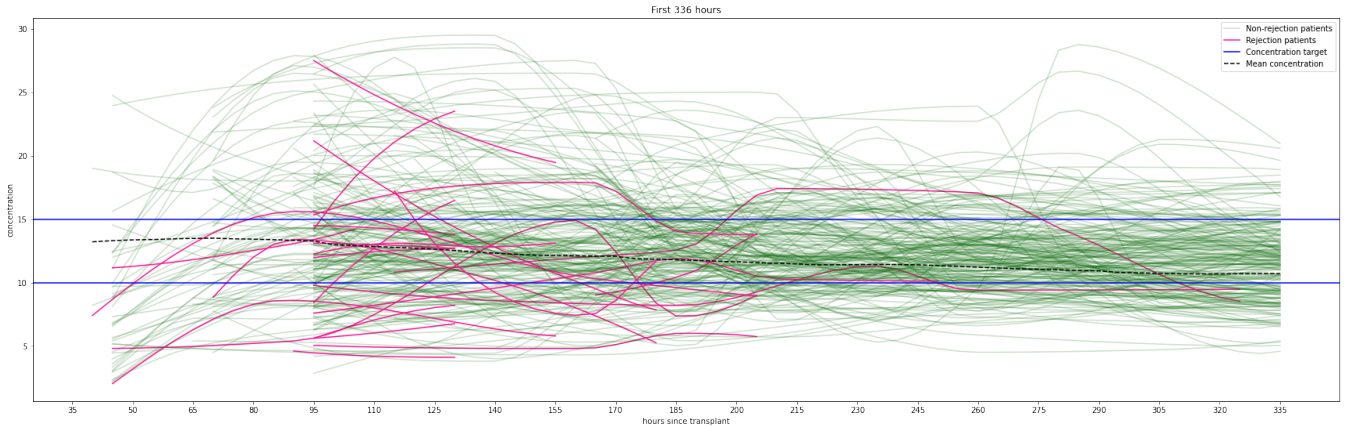


Figure 2: Tacrolimus concentration timeline over first 336 hours after transplant.

of continuous treatments. If the treatment is continuous and there are many treatment levels observed amongst the patients there is not enough data on each treatment level to accurately calculate $P(TreatmentLevel|X)$. In this case treatment being concentration, there are infinite amount of possible values, which makes the use of this treatment variable as it is impractical. The usual solution to this problem is to define levels or buckets of treatment such that more patients are included in each treatment level.

In the trial, target concentrations were defined as: 10-15 ng/mL for the first 15 days, 8-12 ng/mL from day 15 to 28, and 5-10 ng/mL from day 28 onward. Based on these targets, the variable "Hours on Target" (HOT) was created to better understand the relationship between these targets and the success of the treatment. This was achieved by measuring at each time point how many hours the patient had been on target since the start of treatment. Binary information on whether the patient was on target at each time point was also stored for causal analysis.

For time varying analysis, data from dynamic variables was interpolated to generate measurements every 5 hours. The interpolation achieves measurements evenly spaced over time for all patients, a characteristic lacking in the original data and needed for HOT calculations and the time varying causal methods we will use.

At the end of the 3 month mark, 12% of patients rejected their transplanted kidney (this study does not differentiate between biopsy proven and clinically presumed rejections). These rejections happen over the course of the 3 months, as represented in the histogram of Figure 1. This data is used to predict rejection in different ways in our analyses, but one concern remains: in order to predict rejections at any point in time, a model needs examples of patients that rejected their kidney in those times. For example, the models trained in section 5 choose a specific time point and predict rejection from that point onward. For this prediction to be possible there need to be patients in the dataset that did reject their kidneys after the chosen timepoint. These are called positive samples. The amount of positive samples (more specifically, the ratio between positive and negative samples) affects how well a model can classify new cases, since the model needs enough examples of each class to be able to differentiate between them properly. As can be seen in the histogram, the amount of patients that reject at each time point decreases substantially over time, making any predictive analysis difficult after 400 hours. In order to perform analysis on this data, causal and otherwise, this characteristic had to be considered when defining model structures and research questions.

4 Methods

Causal Graph

To gain insight into the intricate relationship between treatment factors and rejection outcomes, we constructed a causal graph. This graph was developed through consultations with the Erasmus MC team of doctors and researchers, who provided insight on the study's design and medical knowledge, as well as their own clinical observations. The aim was to gain a comprehensive understanding of the complexity and confounding factors surrounding treatment and rejection outcomes, and to determine the extent to which tacrolimus concentration is confounded in its effect on the rejection outcome. This understanding informs the structure of the causal analysis, as the nature of the confounders influences what and how to adjust for with causal methods to remove confounding.

The resulting causal graph shown in Figure 3 reveals a complex and interconnected network of causal factors, with multiple variables influencing tacrolimus and each other. Three key variables of interest were identified that might drive treatment outcomes: tacrolimus dosage $\rightarrow$ tacrolimus concentration (the treatment) $\rightarrow$ rejection (the outcome). By looking at the causal graph we can identify two confounders: age of the transplant recipient, and the use of steroid therapy. Older patients suffer from a decrease in the clearance of tacrolimus out of the body, leading to higher concentrations in their bodies. At the same time, younger patients have a stronger immunossupressive system that causes them to reject transplants more easily. This makes age a confounder in the tacrolimus concentration $\rightarrow$ rejection causal

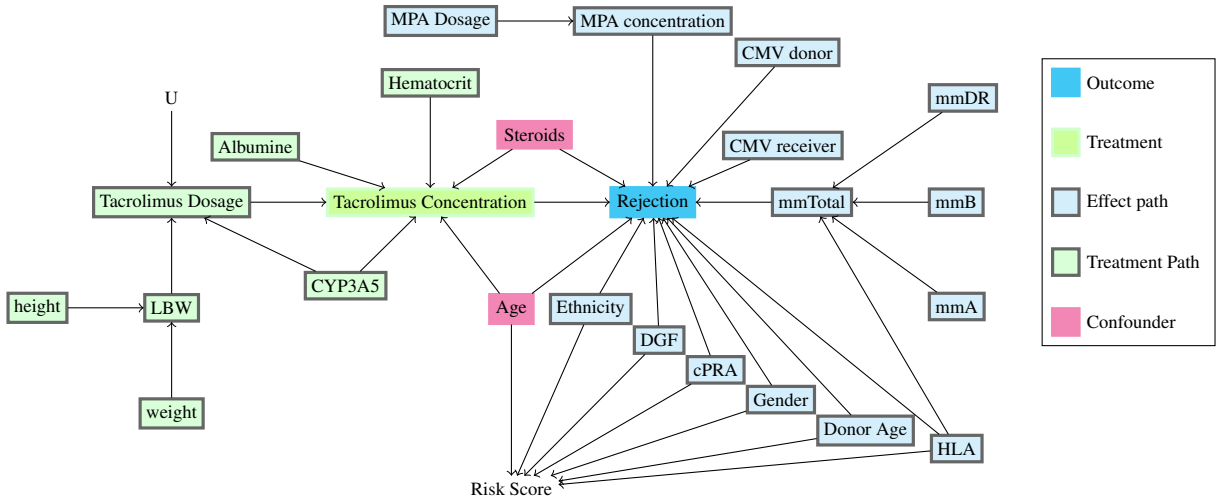


Figure 3: Causal Graph - An arrow between two nodes A $\rightarrow$ B indicates that A causes B. Our study is concerned with the effect of treatment (tacrolimus concentration) on outcome (rejection). A node on the treatment path is then a node belonging to a variable that has a direct (direct arrow) or indirect (through variables that it affects) effect on treatment. Same goes for effect path, where here we include all variables that have a direct or indirect effect on the variable we want to measure, the outcome. The main variables are then the treatment (tacrolimus concentration), the outcome (rejection), and the confounders (in pink). The relationship between these informs the position and relation between all other variables. Abbreviations: MPA - mycophenolic acid, CMV - cytomegalovirus, HLA - Human Leukocyte Antigens, mm - HLA missmatches (type A, B and DR), DGF - delayed graft function, cPRA - Calculated Panel Reactive Antibody, LBW - Lean Body Weight. U stands for Unmeasured factors.

path. Steroid therapy is one of the components of immunosuppressive therapy. These drugs react with tacrolimus, affecting its concentration in the body. At the same time, it is an immunosuppressive treatment, meaning it has an effect on rejection outcome. Once again, this identifies a confounder.

As is common in time-varying treatments, physicians utilize information on past drug concentrations to make decisions on drug dosage. In the event that a patient's concentration is too low, a physician may increase the dosage to avoid rejection. Consequently, all past concentration levels influence current levels, creating treatment-confounder feedback [Hernán and Robins, 2020], where past concentration confounds present concentration and present rejection outcome. This is represented in the time varying causal graph of Figure 4. Adding to this, the rejection outcome can be caused by past tacrolimus levels, as a low concentrations at one time can lead to rejection later on. Both of these facts lead to the conclusion that past treatment is a confounder of present treatment.

The influencing factors on tacrolimus dosage decisions over time introduces the biggest level of uncertainty on the accuracy of the causal graph. While physicians' adjustments are influenced by some measured variables as shown in the graph, there exists a level of freedom and instinct behind these decisions that is unmeasured. In the event that a physician makes a dosing decision based on a variable that influences rejection rate, a confounder is introduced, and if this variable is not included in the list of measured variables, this confounder is unmeasured (U in Figure 3). For the latter situation, the measurement of the causal relationship between tacrolimus concentration and the outcome becomes unidentifiable.

The confounders identified, as well as the complexity and uncertainty of the causal relationships, make the measurement of the causal effect challenging. The amount of data available adds to the challenge, making it unclear whether causal methods will be able to capture the true relationships with confidence. To provide support to the causal methods a predictive analysis is conducted to confirm whether the relationships described in the causal graph are identifiable in the data received.

Predictive Analysis - Modelling choices

In predictive analysis, the choice of models is influenced by the importance of both predictive performance and explainability. In this study, where the primary objective is comparing results across different variables, the focus on explainability is crucial. The selected model should then offer insights into the contribution of each variable to the generated probabilities.

To address this need for explainability, Logistic Regression was initially chosen as the primary model. This selection was based on its inherent capability to offer insights into the impact of individual variables on the resulting probabilities. However, in order to assess whether employing a more complex model could enhance the comparative analysis between models, another model was selected for comparison: XGBoost. Notably, XGBoost represents a tree-based ensemble model, which allows for identification of

the role of specific variables towards the model's output [Chen and Guestrin, 2016].

In the comparison of models, an additional factor of consideration was the choice of evaluation metric. Given that the primary objective centered around predicting the probability of rejection rather than classification, the models were assessed and compared based on the Logloss metric.

The data received used ordinal encoding for categorical variables. Results were compared for models with the original ordinal encoding vs. one hot encoding. Since one hot encoding lead to consistently worse model performance, the original encoding was used. Variables were standardized by removing the mean and scaling to unit variance.

Nested cross-validation was employed to ensure robustness in the predictive models by combining hyperparameter tuning and fitting without duplicating data. Different configurations were tested for number of inner and outer folds. Since the standard error for logloss results of the different configurations did not vary significantly, the combination of 5 folds for outer and 3 folds for inner was chosen. Stratified K Fold (based on the rejection column) was chosen for assigning the folds due to the dataset's significant class imbalance and limited data, allowing for consistent and realistic splits of positive samples throughout the analysis. Hyperparameter tuning addressed two main concerns. To mitigate overfitting, we make sure each leaf requires a higher number of weights for a split, requiring more data in each leaf. Additionally, considering the class imbalance, we made the update step more conservative, as the more intuitive option of using XGBoost's data weighting method was not recommended when needing to obtaining calibrated probabilities [Chen and Guestrin, 2016].

Although calibration of the XGBoost model probabilities was desired to estimate the probability of kidney rejection, a separate model calibration step was not feasible in this dataset due to the need for an additional train/test/validation split. Insufficient data remained after training the predictive model, emphasizing the importance of employing this step in future analyses with larger datasets.

Causal Inference

The primary objective of this study is to understand how tacrolimus concentration levels affect the chance of kidney rejection. The main task is then to measure the causal effect this variable has on the outcome.

Our aim is to determine whether different tacrolimus concentration levels cause the probability of kidney rejection to differ. The intuitive way to calculate this causal relationship is to compare patient outcomes given different levels of treatment. This can be described in the Average Treatment Effect (ATE) as the subtraction of the average outcome (Y) given two different levels of treatment ($A = [a_1, a_2]$): $P(Y|A = a_1) - P(Y|A = a_2)$. This expression can be adapted according to the nature of the treatment or the outcome. One can also plot the distribution of $P(Y|A)$, to better understand how every level of treatment compares. Given the time varying continuous nature of the treatment three levels were defined as average tacrolimus levels below, above and on target. The distribution to calculate is now $P(Rejection|\overline{OnTarget})$. A useful approximation of this, and the distribution ultimately shown in section 6 is $P(Rejection|\%HOT)$.

One might question why we cannot simply construct a predictive model to estimate these probabilities. While predictive analysis can indeed provide probability estimates, it falls short in inferring how these probabilities would change under varying conditions [Pearl, 2009]. This limitation arises from the fact that machine learning models (or any other predictive analysis approach) cannot capture how and if environmental factors are connected within each setting. The joint distribution of tacrolimus and kidney rejections does not inherently provide insights into the avoidance or occurrence of one based on the other. Such information can only be derived from a causal look at the context, as outlined in section 3, through a causal graph [Pearl, 2009].

It is crucial to note that although the data used in this study is derived from a clinical investigation, the study itself did not specifically focus on the effects of tacrolimus concentration. Consequently, the necessary steps to mitigate bias related to tacrolimus (such as randomizing the tacrolimus levels amongst patients such that no outside factors influence tacrolimus concentration) were not implemented. Because of this confounding factors may arise, as indicated in section 4. Thus, while it is feasible to calculate the ATE, there is inherent uncertainty as to whether the results accurately reflect reality without addressing the bias surrounding it.

To mitigate this uncertainty, the confounders that arise from the lack of randomization of the study and the time-varying nature of its variables must be adjusted for. This was done through a method called Marginal Structural Models.

Causal Assumptions In order to perform causal inference, there are a number of assumptions that must be met in order to achieve a valid result. We investigate the validity of our assumptions.

The principle of exchangeability assumes that, following adjustment for confounders, patients who received different levels of treatment would achieve the same outcome as those who actually received those levels of treatment. In our study, for instance, if a patient never achieved the target treatment range within the first 336 hours and experienced kidney rejection, exchangeability implies that another patient who also did reach the target range would have likely experienced rejection as well had they never achieved target. As demonstrated in previous literature [Cole and Hernán, 2008], we rely on the expertise of medical professionals, particularly those involved in the study, to identify and accurately measure the most influential confounding factors, instilling confidence in the appropriateness of our adjustments.

The principle of positivity asserts that every patient has a positive probability of receiving all levels of treatment, indicating that treatment assignment is randomized. Following adjustment, it is essential to ensure that patients can potentially fall into both treated and untreated groups across all confounder levels and timepoints. However, our model may be susceptible to violating the positivity assumption. By transforming the treatment variable (from a continuous concentration measurement to a binary "on target" measure-

ment), the assumption of positivity should be strengthened (reducing the number of levels and the probability of no one receiving treatment in a specific level). Nonetheless, this simplification may lead to a loss of confounder information.

Marginal Structural Models As elucidated in the section 4, the risk of the outcome is impacted by the individual's past exposure history, emphasizing the need to employ appropriate statistical methods to address the confounding introduced by past tacrolimus concentrations. Having recognized the presence of a time-varying confounder in our research, it is essential to account for its influence in order to estimate the causal effect accurately.

Marginal Structural Models (MSMs) have been specifically chosen as the methodology for causal analysis in our study. Prior scholarly discourse has highlighted MSMs as the least biased approach for handling confounders in a time-varying context [Robins et al., 2000]. By leveraging MSMs, we aim to derive robust estimates of the causal effect while effectively adjusting for the confounding influence of the time-varying confounder.

In MSMs, Inverse Probability Weights (IPWs) are utilized to mitigate the confounding effect. This technique creates a pseudo population that emulates the conditions of a randomized clinical trial, where the treatment variable of interest (i.e., whether a patient is within tacrolimus concentration targets) is assigned randomly over time. The weight estimation process is as follows:

$$W_i = \left(\prod_{t=0}^{K} P(A_t = a_{ti} | \bar{A}_{t-1} = \bar{a}_{ti}, \bar{L}_i) \right)^{-1} \quad (1)$$

$$SW_i = \prod_{t=0}^{K} \frac{P(A_t = a_{ti} | \bar{A}_{t-1} = \bar{a}_{ti})}{P(A_t = a_{ti} | \bar{A}_{t-1} = \bar{a}_{ti}, \bar{L}_i)} \quad (2)$$

where A is treatment, L are confounders, K is the length of follow-up W_i is the weight for patient i and SW_i is the stabilized weight for patient i. The lower case a is used here as a concrete value, and $\bar{A}_{t-1}$ stands for treatment history of patient i up until $t-1$ (same goes for $\bar{L}_i$)). Once patients are weighted using either stabilized or normal weights, a predictive model is trained to predict $P(Y|A)$, leading to the MSM.

The difference between the stabilized and normal weights lies in the introduction of the numerator $P(A_t = a_{ti} | \bar{A}_{t-1} = \bar{a}_{ti})$. This numerator is used to stabilize the weights so that the variability in $P(A_t = a_{ti} | \bar{A}_{t-1} = \bar{a}_{ti}, \bar{L}_i)$ can be mitigated. If there is no stabilization the variability in the denominator distribution causes some patients to contribute a very high number of copies of themselves to the weighted predictive analysis, ultimately dominating the analysis [Robins et al., 2000]. Due to the instability of the data and causal relationships described in section 4, stabilized weights were chosen to reduce variability in the resulting models.

Length of treatment As discussed in the section 3, our research employs data derived from a clinical study where observations were made over a period of 3 months. Unlike in [Shahn et al., 2020], described in section 2, the nature of our study does not allow for the use of a period without any censored patients for estimation of treatment effects, since the first rejection happens at t = 114h. This period of time is not representative of the overall relationships during treatment, and the data is scarce, leaving a need o include a larger period of treatment into the analysis. This can only be done after censoring.

Another point to note when choosing when to do our analysis is the difference in causal context. It might not make sense to include treatment at 2 months in the same analysis as treatment at 1 week, as the the causal relationships might not be the same or have the same strength. All of this points towards considering more contained and consistent moments in treatment history.

Patients who experienced kidney rejection were then subsequently censored at the time of rejection, leading to heterogeneity in the number of measurements across patients. This heterogeneity becomes important when employing causal methods, as these necessitate the computation of probabilities throughout the entire follow-up duration. This happens because when the length of follow-up differs among patients, the multiplication involved in Equations 1 and 2 encompasses varying numbers of measurements (different Ks). As these multiplications involve probabilities, a larger number of measurements corresponds to smaller weights. While such bias may be tolerable under negligible differences in follow-up lengths, it should be acknowledged that our study encompasses patients with K values ranging from 128 to 2500 hours. Consequently, this leads to disparate orders of magnitude in the weight calculations, thereby introducing bias in a step that was originally intended to mitigate it.

Avenue of Solution In the realm of survival analysis, the primary goal is to estimate the expected duration until an event occurs. This method finds its application in [Hernán et al., 2000], as described in section 2. The estimation is achieved by assessing the risk of the event at each time point, denoted as t. In contrast to a more classic MSM, where the the outcome model predicts the probability of rejection given the entire treatment history $(P(Rejection|\overline{Tacrolimus}))$, the adapted survival analysis based version focuses on the probability of rejection given treatment history up until a specific timepoint, denoted as $P(Rejection|\overline{Tacrolimus_t})$, where $\overline{Tacrolimus_t}$ denotes the treatment history until time t. This revised approach ensures that, for all uncensored patients at t, an equal amount of measurements from $[start\,of\,measurement : t]$ are utilized in the predictive analysis. Consequently, all patients contribute an identical number of measurements, as the commencement of follow-up is uniform for all individuals, and censored patients at time point t are excluded from consideration.

Other methods were explored (see Appendix). However, due to the uncertainty of this context, a reliable and well supported solution was necessary. Seeing as all other methods were not supported in literature or were not ideal for our context, the use of cutoff of length of treatment as inspired by this method was deemed the best course of action.

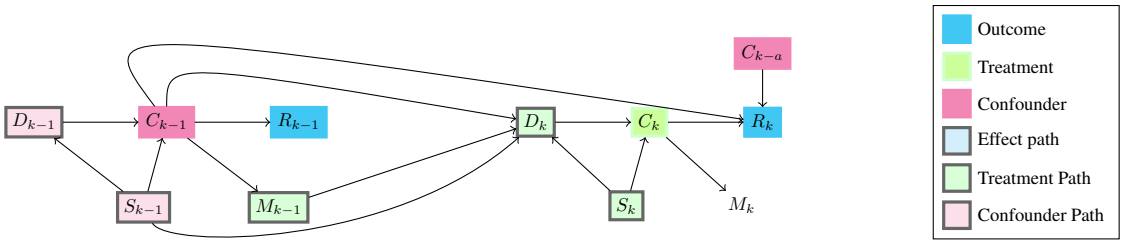


Figure 4: Time Varying Causal Graph - here C_k = tacrolimus concentration at $t = k$, D_t = tacrolimus dosage at $t = k$, R_k = rejection event at $t = k$, M_k = time varying measurements at $t = k$ and S_k = static measurements at $t = k$. Here $a < k$. We can see the treatment confounder feedback that makes C_k a confounder in all future rejection outcomes by looking at the arrows between all Cs and future rejections.

5 Predictive Analysis

Prediction sets

Before causal methods can be applied there is a need to understand the relationships between different variables in the data, and whether tacrolimus has a clear connection to the rejection outcome. To do this predictive models were trained for different sets of variables. Each model received all data of the chosen variables for each patient up until a certain time point, and predicted whether the patient would reject the kidney in the future:

$$P(Rejection_i|\bar{X}_i(t)),$$

where $Rejection_i$ represents a binary variable on whether the patient rejected after t, and $\bar{X}_i(t)$ represents the set of variables until time t for patient i.

By comparing the predictive performance of models with different sets of variables, the study aimed at understanding which variables were more predictive of the outcome. From a more practical point of view, this analysis is useful to inform the medical team on which measurements to watch out for over the course of treatment.

Due to the characteristics of the data, we chose an early point in time (t = 336h, the first 14 days after transplant) as the input to predict rejection outcomes. Several sets of variables were defined, and Logistic Regression and XGBoost models were trained for each set. For each time varying variable the mean, standard deviation, 25th, 50th, and 75th percentiles, as well as the min and max values up until that point, were calculated.

One of the features examined by healthcare professionals before a kidney transplant is the risk score, which is the result of the sum of risk scores of six other features: age of donor, age of patient, ethnicity, panel of reactive antibodies (PRA), delayed graft function, and HLA mismatches. This feature is a summary of several features thought to be indicators of risk of transplant rejection.

The second set of variables was related to tacrolimus. Several variables were used to investigate the relationship between tacrolimus and the outcome: tacrolimus concentration, tacrolimus dosage, HOT and age. Age was added since it was identified as a confounder of the effect of tacrolimus on rejection.

Furthermore, three additional models were developed to augment the analysis. The first model exclusively incorporated the static variables, while the second model solely focused on the dynamic variables. The final model encompassed all variables, both static and dynamic. By segregating the models in this manner, we aimed to elucidate the influence of treatment-related variables and the body's response (captured by dynamic variables) on the predictive performance of the model. It is expected that the inclusion of dynamic variables would enhance the model's predictive capabilities, as it would indicate that these variables offer pertinent information regarding the outcome.

Results The results for the predictive analysis are shown in Table 1. Ultimately, the resemblance between variable set results and the high standard deviation between folds makes it difficult to draw definitive conclusions on which set seems to be the most predictive. This leads us to state that these results should not be used in a clinical setting, but some insights can be obtained comparatively.

The objective was to spot a trend across models in terms of comparative results between sets of variables by ordering the logloss results between variable sets. Unfortunately, that is not possible in this case, since both models produced a different order of sorted logloss results between variable sets. The common factor between models lies in the fact that the complete variable set performs equally or better than static and dynamic in both models. This tells us that there is some benefit in combining the dynamic and static variables.

Since the predictive power of tacrolimus is not clear in this analysis, as well as the benefits of the use of dynamic variables in general, we look to the feature selection results for clarification.

Feature Selection

To gain deeper insights into the predictive analysis, individual examination of the dynamic variables was undertaken. A manual and adapted version of forward feature selection was employed. The static model served as the baseline against which the incremental contributions of each dynamic variable were evaluated. The selection criterion was based on the reduction in average log loss over nested cross validation folds, enabling the identification of the variable that yielded the most substantial enhancement in predictive performance.

Table 1: Predictive analysis results. Here we compare logloss and ROC AUC for different sets of variables. The complete set (All) is consistently better than both static and dynamic sets, leading us to believe that the use of both is best for rejection prediction.

	Logloss +/- std		ROC AUC +/- std	
	Log. Reg.	XGBoost	Log. Reg.	XGBoost
Risk Score	0.179 +/- 0.030	0.177 +/- 0.005	0.641 +/- 0.122	0.562 +/- 0.115
Tacrolimus	0.177 +/- 0.027	0.181 +/- 0.001	0.412 +/- 0.159	0.436 +/- 0.113
Static	0.180 +/- 0.033	0.177 +/- 0.009	0.738 +/- 0.124	0.575 +/- 0.109
Dynamic	0.176 +/- 0.029	0.179 +/- 0.001	0.639 +/- 0.203	0.502 +/- 0.064
All	0.176 +/- 0.029	0.177 +/- 0.002	0.658 +/- 0.116	0.538 +/- 0.052

Subsequently, the selected variable was incorporated into the model, and the iterative process continued until no further improvement was observed.

For this analysis, two distinct sets of variables were included with the aim of identifying potential indicators of rejection during treatment, and more specifically the relationship between tacrolimus and the outcome. The characterization of time-varying variables followed the same methodology employed in the predictive analysis. The initial set comprised the entire collection of dynamic variables, and subsequently, variables specifically related to tacrolimus were selected. The objective was to unveil the crucial correlations that would guide a prediction model in leveraging specific variables for accurate predictions. This process is partially inspired by the pharmacokinetic method stepwise covariate modelling, where the chosen covariates are tested one by one for their statistical significance based on their obtained objective function value [Mould and Upton, 2013].

Results In Table 2 we look at the variables selected for models at t = 336h and the order of their selection, once again trying to reach some consensus between models. If a variable is selected first (in the static + 1 and static + 1 tac rows), then it is the variable amongst it set to improve the predictive performance of the static variables the most. The same logic is then applied for next rows. Once again the high standard deviation between folds makes it difficult to draw definitive conclusions, but there are some common results to point out.

Since the objective was to identify the variables are comparatively better and most inform the rejection prediction, a choice was made to stop once specific variables were no longer standing out. The fourth and fifth round was performed for both Logistic Regression and XGBoost (see Appendix) but since no single variables stood out a choice was made to stop. A different timepoint (t = 168h) was tested, with much more certainty in the outcomes (see Appendix). For t = 336h, there is in some cases no one variable that stands out in terms of predictive performance, leading to a confusing comparative analysis.

The top two dynamic variables are similar in both methods. Something to note here is the difference between a treatment (MPA) and a clinical measurement (creatinine), which will be explored further in the discussion. The second insight is the fact that tacrolimus does not stand out particularly stand out amongst dynamic variables, and the results of the tacrolimus related set are not consistent between models. This makes the association between tacrolimus concentration and outcome very uncertain. A more specific measurement of this relationship might be possible through causal analysis.

6 Causal Analysis

As mentioned in the section 4, a MSM is made up of two parts: the calculation of IPWs and the mean outcome model.

Inverse Probability Weights

As was highlighted in section 4, there was a need to address the bias created by the disparity in follow up lengths. Taking from [Hernán et al., 2000], discussed in section 2, a specific timepoint was chosen for censoring patients when calculating the inverse probability weights (IPWs). At timepoint t = 336 hours post-transplantation, which marks the completion of the initial concentration target, patients who had not experienced rejection events were censored. This means we only consider data from before t = 336h only for patients with rejection timestamps after 336h.

To ensure stable probabilities and minimize variance, a total of nine equally spaced timepoints were sampled within the interval [95, 336]. This range was selected since all patients are guaranteed to have at least one measurement by t = 95 hours.

Considering the unique characteristics of tacrolimus concentration measurements mentioned in section 3, a simplified measurement approach was adopted. In the context of Equation 2, the treatment $A_{t,i}$ was represented by the binary variable $OnTarget_{t,i}$, indicating whether patient i was within the target dosage range for tacrolimus at time t. The target history $\overline{A_{t,i}}$ was defined as a set of variables including the value of $\overline{OnTarget_{t,i}}$ at the previous timepoint, as well as the mean and standard deviation of previous on-target measurements up to the current timepoint.

Data was collected and consolidated for each timepoint within the specified interval. Logistic Regression models were trained separately for both the numerator and denominator probability distributions in Equation 2. The fitted models were utilized to calculate the probability $P(OnTarget_{t,i} = ontarget_{t,i})$ for each patient i, and subsequently, the IPWs were computed by dividing the numerator and denominator probabilities. After obtaining the

Table 2: Feature Selection analysis results at t = 336h. There is a pattern in the top two variables in the first set, and standard deviation of creatinine and the MPA dose seem to be correlated to the rejection probability.

	Logistic Regression		XGBoost	
	best variable	logloss	best variable	logloss
static + 1 variable	$creat\ std$	0.164 +/- 0.042	$mpa\_dose\ mean$	0.174 +/- 0.006
static + 2 variable	$mpa\_dose\ std$	0.160 +/- 0.034	$creat\ std$	0.173 +/- 0.002
static + 3 variable	$hct\ max, creat\ max$	0.158 +/- 0.039, 0.035	$tac\_con\ mean, tac\_dose\ mean, hctmean, albmean, creatmin, tac\_conmax, tac\_dosemax, albmax$	0.172 +/- 0.00(1,3,4)
static + 1 tac variable	$HOT\ mean, HOT\ min, tac\_con\ std$	0.167 +/- 0.035	$tac\_dose\ max$	0.176 +/- 0.004
static + 2 tac variable	$none$		$none$	

the denominator and numerator probability distributions, the following calculation was performed:

$$\exp\left(\sum_{t=times}^{K}\log\left(\frac{P(OnTarget_t|\overline{OnTarget}_t)}{P(OnTarget_t|\overline{OnTarget}_t,\bar{L}_t)}\right)\right)$$
$$, times = [95, 125, 155, 185, 215, 245, 275, 305, 335] \quad (3)$$

Here $K = 336$ for all patients as mentioned above. We are considering 9 weights only (by generating the probabilities once every 30 hours) in order to reduce variance in the results. When compared to Equation 2 we chose to calculate the exponent of the sum of the logarithms of the probabilities instead of the multiplication of the probabilities to aid the calculations with small values.

The ultimate objective was to obtain stable weights without excessive variance that would impede their utilization in prediction models. This variance is mainly caused by a disagreement in predictions by the numerator and denominator predictors, and is exacerbated when any of these predictors is particularly confident about their prediction. In order to achieve the most well-behaved weights several experiments were performed. These are described in the Appendix . Ultimately, only the established confounders were included in $\bar{L}_t$, and no regularization was applied to the estimation of the probabilities.

Outcome Model

Following the computation of the weights, a univariate Logistic Regression based on the variable $\%HOT$ was weighted and trained to predict rejection $P(Rejection_i|\%HOT)$ to better analyse the effect of treatment on outcome. This procedure was repeated on 500 bootstrap samples. Data was sampled from patients still uncensored at t = 336h and sampling was stratified according to the outcome. The results are displayed in the next section.

A model was also trained to predict the probability $P(Rejection_i|\bar{X}_{t,i})$ using the calculated weights. A Logistic Regression as well as an XGBoost model were trained on a selection of variables based on the results of the Feature Selection. The training procedure used was the same as for the predictive analysis. The primary goal was to obtain accurate probability estimations, enabling the determination of treatment levels for the calculation of the average treatment effect (ATE) of tacrolimus on rejection.

Both univariate and full models were weighted during training, resulting in full Marginal Structural Models.

Addtionally, a model was trained for the effect of %HOT under different ages (see Appendix). The objective was to understand how the treatment effect changes under different demographic groups. This sets up possibilities for future analyses.

Results

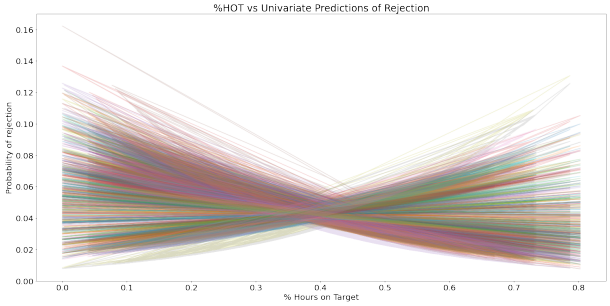


Figure 5: MSM results - variation of rejection probability over percentage of HOT over 500 samples

The results displayed in Figure 5 display the probabilities obtained from the 500 data samples. There is some variance in terms of direction of the curve. To better understand the trend in terms of positive or negative effect, the coefficient for $\%HOT$ is displayed in Figure 6. Here we can see that the distribution is centered below zero. A negative coefficient tells us that with higher $\%HOT$ comes a lower probability of rejection. Since these are our adjusted estimates, we can say that on average tacrolimus levels between 10-15 ng/mL for the first 15 days cause the rejection probability to go down.

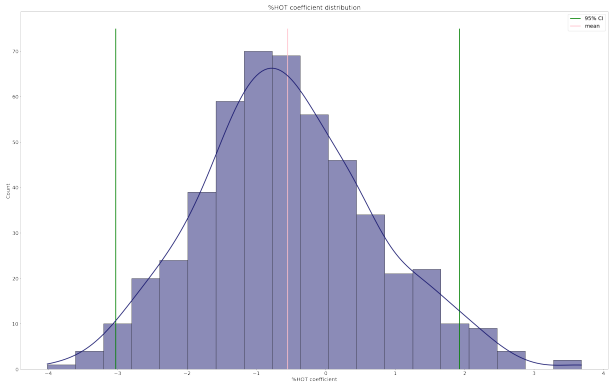


Figure 6: %HOT coefficient distribution over 500 runs. The mean is negative, indicating a negative causal relationship between %HOT and rejection probability

7 Discussion

Throughout our analyses we have found a significant amount of variation in our results. We will now consider the possible reasons behind this variation, and what can be done in future studies to prevent it.

Causal Graph Uncertainty in the causal graph used in this study raises important considerations regarding its influence on the final results. The complexity of measuring causal effects is compounded by the challenge of quantifying physician decisions, which may play a crucial role in the ultimate outcome but are not easily measurable. Furthermore, the dynamic nature of these decisions adds an additional layer of complexity that is difficult to capture accurately.

Certain concepts pose challenges in representing them within a causal graph. For instance, the notion of exposure consistency, referring to variance of drug concentration over time, seems to be a possible cause for rejection, more than just concentration levels, but it is challenging to incorporate in a causal graph and represent it in the causal context. The same happens for the time since transplantation: the fact that a patient has made it past the first two weeks influences their chances of rejection, but time itself is difficult to represent as a causal factor.

The lack of consistency of causal relationships over time further exacerbates the complexity of the context. The treatment context at 48 hours differs significantly from that at 1000 hours, and attempting to encompass both contexts within a single analysis leads to general results that might not be realistic. Separating the analysis would, however, be challenging due to the limited amount of available data.

Complexity vs. Explainability While explainability was prioritized by choosing models such as XGBoost that allow for tracking and interpreting variable importance, it may not necessarily address the core challenges of the problem at hand, as is evident in the results. The added complexity of XGBoost is evidenced in the smaller deviations in results. This highlights the importance of balance between model complexity and interpretability, particularly in the medical domain where practical applicability is essential.

Introducing more complexity in the appropriate direction could potentially yield more stable results, providing a more comprehensive understanding of the problem and facilitating a clearer and consistent comparison of different sets of variables. This however is done at the potential cost of reduced explainability. Striking the right balance between complexity and interpretability remains a challenge that requires careful consideration.

Predictive Analysis Regarding the predictive analysis conducted in this study, different types of variables were identified as being predictive, as well as different timepoints. Variables such as creatinine that emerged as predictive from the feature selection analysis, can be compared to treatment variables such as MPA dose. Creatinine, for example, serves as an informative indicator but does not directly influence the probability of rejection. Conversely, treatments such as MPA or tacrolimus not only exhibits a correlation but also plays a causal role in altering outcomes, thereby providing a viable target for intervention. Causal inference methods are essential in unraveling these intricate relationships and informing decision-making.

Additional feature selection experiments were conducted with a different timepoint in mind (see Appendix). Here hematocrit (hct) appeared as a much clearer favorite in comparison to the results at t = 336h. Since the predictiveness of hct is tied to the time since surgery, this results shows that taking account context choosing the correct timepoint is important to the analysis, as one must make sure that it is representative of what is being studied. In this case the first days after surgery are not representative of general treatment, hence the models will identify different feature relationships.

Notwithstanding the suboptimal results of the predictive analysis, it is noteworthy that there is minimal variation observed between the Logistic Regression and XGBoost models. This outcome is unexpected since these models are fundamentally distinct in terms of linearity, probabilistic nature, and complexity. One would anticipate that these divergent characteristics would capture different levels of intricacies within the data, resulting in disparate accuracies. However, such differentiation is not apparent in the findings. This observation raises three potential explanations that warrant exploration in future studies.

Firstly, it is conceivable that the data exhibits insufficient correlation to establish meaningful associations and predict rejection events. However, this possibility seems unlikely, given that medical practitioners have successfully utilized some of these metrics for years in guiding treatments with positive outcomes.

Secondly, the observed findings may be influenced by the nature of the clinical study and the characteristics of the study population. The stringent control measures and minimal treatment variations in the study aim to prevent rejections, resulting in a lack of instances where treatment levels could induce rejection. This absence of data reflecting covariate levels associated with rejection events could contribute to the similarities observed between models. Related to this, the study population exclusively consists of patients

compatible with their kidney transplants, eliminating values of static variables that could potentially contribute to rejections from the dataset. Consequently, the absence of examples depicting covariate levels linked to rejection may be attributed to the deliberate control measures employed to avoid such averse events, implying that rejections in this study might be influenced by latent factors not captured by the covariates.

Finally, the limited number of rejection instances within the dataset hampers the model's ability to discern distinct patterns.

Besides the confidence in the results, the lack of consistency in terms of sorting of variable sets and feature selection variables must also be discussed. As mentioned early, Logistic Regression and XGBoost differ in many essential traits. These differences could be enough to justify the differences in results, although more significant differences in performance would be expected.

Comparison to similar study The comparative study [Truchot et al., 2022] mentioned in section 2 is worth discussing. Although our study differs in prediction objectives (outcome prediction vs. survival analysis), data availability, and follow-up periods, there are noteworthy points of overlap that can contribute to the interpretation of our findings.

There are notable similarities between the two studies, including an emphasis on interpretability, the utilization of XGBoost, and some shared findings. Both studies explored the interpretability aspect by examining individual covariate performance, either through true feature importance or model performance. This underscores the validity of our chosen approach, independent of specific outcomes. The adoption of XGBoost in both studies highlights its logical relevance as a model choice, despite sub optimal calibration observed in the referenced study. Logistic Regression, with its inherent probability prediction, would theoretically yield superior calibration, although testing this hypothesis was not feasible in either study. Notably, the referenced study found no significant performance differences among the tested models, aligning with our own findings in a distinct context.

Importance of Causal Inference The example of the amount of methods used in [Truchot et al., 2022] underscore the fundamental distinction between predictive analysis and causal analysis. Machine learning techniques have made significant strides in predictive power, and although the methods employed in this study may not have yielded optimal predictive performance, alternative approaches exist. In contrast, causal inference methods have not yet reached a level of research and application that allows for straightforward solutions in complex scenarios such as the one presented here. Extensive research and adaptations were necessary to align the available methods with the data and problem setting, yet confidence in the results remains limited given the complexities and paucity of data.

However, it is essential to emphasize the value of advancing causal inference methods. When physicians inquire about the effect of tacrolimus concentration on patient rejections and the specific concentrations required to mitigate rejection risks, understanding the causal relationship between the cause (tacrolimus) and the effect (rejection) becomes paramount. No other methodology can provide definitive answers to these critical questions. Hence, investing in the development of causal inference methods not only pushes the boundaries of machine learning but also enables a deeper understanding of the underlying processes inherent in the data, ultimately yielding more informative and actionable insights.

Future model choices Alternative causal analysis methods exist alongside MSMs, prompting the need to justify their selection and assess their appropriateness. MSMs were chosen for their interpretability, which sets them apart from other approaches and enables the extension of machine learning models for causal analysis [Hernán et al., 2000]. This flexibility in model choice and customization is advantageous in contexts where interpretability of both results and model decisions is crucial for a medical team. MSMs provide visibility into model coefficients or importance, and their final plot effectively portrays the trend and magnitude of the treatment-outcome relationship.

Another modelling choice made was in the adaptation of the survival analysis MSM model [Hernán et al., 2000]. Because of the nature of our data it was not possible to calculate $P(Rejection_t)$ as the paper does. Another reason for this choice is the inherent assumption of proportional hazards in survival analysis. As has been mentioned, this study takes data from different periods of treatment with different causal contexts. However, for contained time periods, this approach might lead to more significant causal conclusions. This happens since the data up until time t is more closely related to the prediction at time t than any other. Hence, the causal relationships obtained through the MSM would also be more logical.

However, confidence in the results poses a major limitation. Other methods, like Structural Nested Models, hold promise for potentially narrower confidence intervals [Hernán et al., 2000]. Four key factors—causal models, predictive models, data, and confounder specification—must be considered when making future modeling choices. Future research should prioritize addressing data-related challenges, as they underpin various issues encountered. As discussed earlier, class imbalance and limited data undermine confidence in both predictive and causal methods. Moreover, comparison with other studies (and within ours) reveals that different models yield comparable outcomes. Consequently, improving data quality can potentially yield better results while maintaining the interpretability offered by current methods.

Data Necessities for future research Since the results obtained were constrained by the nature of the data used, we leave suggestions for future research on this topic.

The significant variation in rejection densities across different time periods during the analysis is one of the main problems. The data requirements for achieving reliable results are contingent on the class imbalance observed within each temporal section. Consequently, future investigations should concentrate on periods characterized by consistent

rejection incidence to enhance predictive accuracy.

To enable feasible causal inference analyses, establishing a follow-up period with equal measurements for all patients is crucial. For example, gathering 200-500 uncensored patients at week 3 allows general predictions of rejection. When this is not possible, scatter instances of rejection throughout the follow-up duration enhance the ability to predict rejection probability at different time points (t), denoted as $P(Rejection_i(t)|X_i)$. The objective is not to artificially inflate rejection incidence but to reflect the expected occurrence rate. Mitigating class imbalance requires additional data to augment positive samples for analysis.

There are several other factors that future studies should aim to address that could contribute to improved results. Firstly, it is highly plausible that important variables are missing from the current analysis, which could potentially enhance the prediction of rejection events, such as steroid treatment, which may serve as a confounding factor and exert a significant impact on the outcomes. Including this variable in future analyses could provide valuable insights.

Furthermore, the consistency of measurements warrants consideration. In this study, measurements were interpolated to establish a uniform measurement schedule. However, given that patients progress through different treatment stages, the availability of measurements varies. It is unrealistic to expect identical measurement intervals for patients receiving treatment at home compared to those in interim care. Consequently, a more viable approach would be to treat these distinct stages as separate analyses, ensuring that each stage includes an adequate representation of rejection events by gathering a sufficient amount of data.

By addressing these factors in future studies, a more comprehensive understanding of the variables influencing rejection events can be achieved, leading to improved prediction models and enhanced clinical insights.

8 Conclusion

In this study, we have outlined the methodology for investigating the causal effect of tacrolimus concentration on the probability of kidney rejection based on the available data. The results obtained from both the predictive and causal analyses demonstrated notable variability; however, a trend emerged from the causal analysis indicating that higher levels of percentage of hours on target ($\%HOT$) were associated with reduced chances of rejection. These findings underscore the continued relevance and significance of employing causal analysis to uncover the relationship of interest.

To further advance our understanding, future research endeavors should prioritize the analysis of alternative datasets, specifically those encompassing more constrained follow-up periods, larger sample sizes, and, if realistic, a diminished class imbalance. Such investigations hold the potential to enhance the robustness and generalizability of our findings, ultimately leading to more accurate predictions and improved patient outcomes in the context of kidney transplantation.

9 Acknowledgments

The author would like to express their sincere gratitude to Jesse H. Krijthe, for their invaluable guidance and support throughout the course of this research. Thanks are also due to Maaike Schagen and the Erasmus MC Transplant Institute for their insights and cooperation. Finally, the author would like to thank family and friends for their encouragement and helpful discussions during the preparation of this paper. Their input has been instrumental in shaping our ideas and improving the clarity of the work.

References

[Bouamar et al., 2013] Bouamar, R., Shuker, N., Hesselink, D. A., Weimar, W., Ekberg, H., Kaplan, B., Bernasconi, C., and van Gelder, T. (2013). Tacrolimus predose concentrations do not predict the risk of acute rejection after renal transplantation: a pooled analysis from three randomized-controlled clinical trials(†). *Am J Transplant*, 13(5):1253–1261.

[Brunet et al., 2019] Brunet, M., van Gelder, T., Åsberg, A., Haufroid, V., Hesselink, D. A., Langman, L., Lemaitre, F., Marquet, P., Seger, C., Shipkova, M., Vinks, A., Wallemacq, P., Wieland, E., Woillard, J. B., Barten, M. J., Budde, K., Colom, H., Dieterlen, M.-T., Elens, L., Johnson-Davis, K. L., Kunicki, P. K., MacPhee, I., Masuda, S., Mathew, B. S., Millán, O., Mizuno, T., Moes, D.-J. A. R., Monchaud, C., Noceti, O., Pawinski, T., Picard, N., van Schaik, R., Sommerer, C., Vethe, N. T., de Winter, B., Christians, U., and Bergan, S. (2019). Therapeutic drug monitoring of Tacrolimus-Personalized therapy: Second consensus report. *Ther Drug Monit*, 41(3):261–307.

[Buckley et al., 2015] Buckley, J. P., Keil, A. P., McGrath, L. J., and Edwards, J. K. (2015). Evolving methods for inference in the presence of healthy worker survivor bias. *Epidemiology*, 26(2):204–212.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

[Cole and Hernán, 2008] Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*, 168(6):656–664.

[Flanders et al., 1993] Flanders, W. D., Cárdenas, V. M., and Austin, H. (1993). Confounding by time since hire in internal comparisons of cumulative exposure in occupational cohort studies. *Epidemiology*, 4(4):336–341.

[Fox and Collier, 1977] Fox, A. J. and Collier, P. F. (1977). Mortality experience of workers exposed to vinyl chloride monomer in the manufacture of polyvinyl chloride in great britain. *Br J Ind Med*, 34(1):1–10.

[Gatault et al., 2017] Gatault, P., Kamar, N., Büchler, M., Colosio, C., Bertrand, D., Durrbach, A., Albano, L., Rivalan, J., Le Meur, Y., Essig, M., Bouvier, N., Legendre, C., Moulin, B., Heng, A.-E., Weestel, P.-F.,

Sayegh, J., Charpentier, B., Rostaing, L., Thervet, E., and Lebranchu, Y. (2017). Reduction of Extended-Release tacrolimus dose in Low-Immunological-Risk kidney transplant recipients increases risk of rejection and appearance of Donor-Specific antibodies: A randomized study. *Am J Transplant*, 17(5):1370–1379.

[Hernán, 2018] Hernán, M. A. (2018). How to estimate the effect of treatment duration on survival outcomes using observational data. *BMJ*, 360.

[Hernán et al., 2000] Hernán, M. A., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570.

[Hernán and Robins, 2020] Hernán, M. and Robins, J. (2020). *Causal Inference: What If*. CRC Press.

[Hooks, 1994] Hooks, M. A. (1994). Tacrolimus, a new immunosuppressant—a review of the literature. *Annals of Pharmacotherapy*, 28(4):501–511. PMID: 7518710.

[Israni et al., 2013] Israni, A. K., Riad, S. M., Leduc, R., Oetting, W. S., Guan, W., Schladt, D., Matas, A. J., Jacobson, P. A., and DeKAF Genomics Investigators (2013). Tacrolimus trough levels after month 3 as a predictor of acute rejection following kidney transplantation: a lesson learned from DeKAF genomics. *Transpl Int*, 26(10):982–989.

[Loupy et al., 2019] Loupy, A., Aubert, O., Orandi, B. J., Naesens, M., Bouatou, Y., Raynaud, M., Divard, G., Jackson, A. M., Viglietti, D., Giral, M., Kamar, N., Thaunat, O., Morelon, E., Delahousse, M., Kuypers, D., Hertig, A., Rondeau, E., Bailly, E., Eskandary, F., Böhmig, G., Gupta, G., Glotz, D., Legendre, C., Montgomery, R. A., Stegall, M. D., Empana, J.-P., Jouven, X., Segev, D. L., and Lefaucheur, C. (2019). Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study. *BMJ*, 366.

[Mould and Upton, 2013] Mould, D. R. and Upton, R. N. (2013). Basic concepts in population modeling, simulation, and model-based drug development-part 2: introduction to pharmacokinetic modeling methods. *CPT Pharmacometrics Syst Pharmacol*, 2(4):e38.

[Pearl, 2009] Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146.

[Robins et al., 2000] Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

[Sapir-Pichhadze et al., 2014] Sapir-Pichhadze, R., Wang, Y., Famure, O., Li, Y., and Kim, S. J. (2014). Time-dependent variability in tacrolimus trough blood levels is a risk factor for late kidney transplant failure. *Kidney International*, 85(6):1404–1411.

[Shahn et al., 2020] Shahn, Z., Shapiro, N. I., Tyler, P. D., Talmor, D., and Lehman, L.-W. H. (2020). Fluid-limiting treatment strategies among sepsis patients in the ICU: a retrospective causal analysis. *Crit Care*, 24(1):62.

[Shuker et al., 2016] Shuker, N., Bouamar, R., van Schaik, R., Clahsen-van Groningen, M., Damman, J., Baan, C., van de Wetering, J., Rowshani, A., Weimar, W., van Gelder, T., and Hesselink, D. (2016). A randomized controlled trial comparing the efficacy of cyp3a5 genotype-based with body-weight-based tacrolimus dosing after living donor kidney transplantation. *American Journal of Transplantation*, 16(7):2085–2096.

[Truchot et al., 2022] Truchot, A., Raynaud, M., Kamar, N., Naesens, M., Legendre, C., Delahousse, M., Thaunat, O., Buchler, M., Crespo, M., Linhares, K., Orandi, B. J., Akalin, E., Pujol, G. S., Silva, Jr, H. T., Gupta, G., Segev, D. L., Jouven, X., Bentall, A. J., Stegall, M. D., Lefaucheur, C., Aubert, O., and Loupy, A. (2022). Machine learning does not outperform traditional statistical modelling for kidney allograft failure prediction. *Kidney Int*, 103(5):936–948.

A Appendix

Length of treatment Solutions considered

Changes in interpolation approach Another potential solution to the bias caused by the inconstant number of measurements between patients with different lengths of follow-up is to reconsider the amount of measurements used for each patient. Currently, patients are measured at irregular time points, and these measurements are interpolated to create uniform 5-hour intervals between each measurement for each patient. However, this interpolation method gives rise to the aforementioned bias.

An alternative approach to mitigate this bias is to modify the interpolation method to ensure that all patients have the same number of measurements. This would ensure that when computing IP weights, all patients have an equal number of elements in the multiplication, eliminating the bias. However, this approach would entail measuring variables at different time points for different patients. For example, if each patient is measured 20 times and patient A experiences rejection at $t = 200h$ while patient B experiences rejection at $t = 2000h$, then patient A's third measurement would occur at $t = 30h$, while patient B's would occur at $t = 300h$.

It is, however, critical to ensure that the time intervals are not too far apart such that the causal relationship between treatment (i.e., tacrolimus concentration) at time t and $t + 1$ holds. If the time intervals are too distant, the assumption that healthcare professionals use previous treatment information to determine present treatment may no longer hold. Unfortunately due to the significant differences in lengths of follow-up this would indeed happen, as for patients with longer follow-up periods this would lead to interpolation periods such that from one time point to another there would no longer be any causal relationship (the tacrolimus administered at one point would no longer exist in the body in the next time point, and the doctor would not use the information on the last time point to make dosing decisions in the current time point). Because of this this decision is not applicable in the current context.

Smoothing The problem under consideration involves a bias resulting from an uneven number of measurements

across patients with different lengths of follow-up. This bias stems from the multiplication of varying numbers of terms, leading to different results. A possible solution is to employ the geometric mean to smooth out the multiplication, thereby rendering the number of terms irrelevant while preserving all the pertinent information. Specifically, the formula for the geometric mean is:

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

Implementation of this approach would entail taking the K_i-th root, where K_i is the length of follow-up for patient i.

To the best of our knowledge, this solution has not been previously explored in the literature. Consequently, it requires rigorous mathematical proof and a thorough analysis to ascertain the correctness and significance of the results. Our initial conjecture is that this added calculation affects all patients in a similar fashion and the probabilities derived from IPW remain unchanged, thereby preserving the outcome's integrity. However, one potential source of concern is the potential loss of information incurred during the smoothing process.

Adjusting for bias The presence of bias introduced by the length of follow-up in the analysis was a critical consideration. One approach to address this issue is to treat it as a confounding factor and explore potential methods for its adjustment.

Previous investigations into the impact of follow-up duration on bias have led to studies examining the concept of healthy worker bias and strategies for its mitigation. These studies propose that if treatment effects only manifest after a certain period of time, there is a bias associated with the duration of treatment exposure. For instance, in [Flanders et al., 1993], the time since hire is recognized as a confounder as it influences both cumulative treatment exposure and mortality outcomes. When the duration since hire is short, there may be insufficient exposure to yield mortality. This related to the present study since the length of follow-up (here time since hire) is being acknowledged as bias, and efforts are being made to diminish it. However, the suggested solution is to focus solely on time periods wherein the relationship between treatment and outcome is observed, effectively excluding the initial period during which treatment cannot affect the outcome [Fox and Collier, 1977]. This approach introduces potential selection bias as it exclusively considers healthy individuals beyond a certain time point [Buckley et al., 2015]. This solution is not an option in the present study seeing as considering only patients amongst certain lengths of follow up would lead to a level of class imbalance that would make any prediction of the rejection outcome impossible with the available data.

Alternatively, direct adjustment for this bias is another viable option. In [Hernán, 2018], the authors describe the bias that arises when patients with varying follow-up times are included in the same study. The authors investigate the causal effect of treatment duration on treatment outcomes, offering valuable insights. This study chooses to quantify the effect that different lengths of treatment have on the outcome. While this is useful for understanding the relationship length of treatment has on the data, our objective is to understand the effect of treatment, so calculating the effect of length of treatment would not work towards that goal.

Supplementary Results - Feature Selection

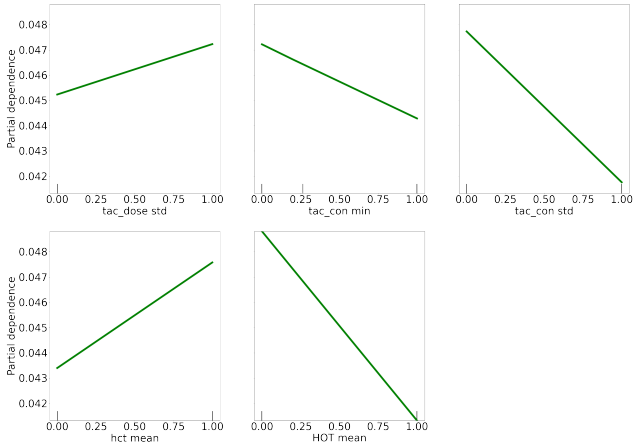


Figure 7: Partial Dependence Plots for some of the top variables (from feature selection). Here a Logistic Regression model was trained with the static set + the variables selected from the feature selection

For t = 168h The variable `hct mean` is consistently the variable that contributes the most predictive power to the static variables dataset. After talks with the medical team, the beginning of an interpretation of this result was obtained. Hematocrit (hct) is the percentage by volume of red cells in the patient's blood. This means it is directly tied to drug concentrations, as represented in Figure 3 (the causal graph). After the transplant operation, part of the recovery process involves achieving more stable (higher) levels of hct. This means healthier patients will display higher levels of hct, which means there will be more red blood cells to measure in the blood. Since tacrolimus binds to red blood cells, higher concentrations will be measured. Since lower concentrations of tacrolimus are tied to rejection, the high values of hct should be tied to lower probabilities of rejection. This is indeed observed. One detail to consider is that the correlation of hct to rejection goes through tacrolimus concentration. This should mean that this effect should be present through tacrolimus concentration first and residually through hct, which was not observed (`hct mean` was consistently picked first, and its effect is present even in models where both measurements are present, as evidenced by Figure 7)

The second finding is the consistency in the second pick. Although not the same variable was picked in second place in both models, both variables picked were an expression of tacrolimus. Once again the deviation in results is considerable, so there is no way to be certain on the validity of the results. To better understand which tacrolimus related vari-

Table 3: Feature Selection analysis results at t = 168h

	Logistic Regression		XGBoost	
	best variable	logloss	best variable	logloss
static + 1 variable	$hctmean$	0.274 +/- 0.078	$hctmean$	0.273 +/- 0.074
static + 2 variable	tac_conmin	0.270 +/- 0.074	$tac\_dosestd$	0.272 +/- 0.075
static + 3 variable	$hctmax$	0.267 +/- 0.073	mpa_conmin	0.271 +/- 0.074
static + 1 tac variable	tac_conmin	0.274 +/- 0.080	tac_constd	0.274 +/- 0.088
static + 2 tac variable	$HOTmin$	0.272 +/- 0.079	$tac\_dosemax$	0.271 +/- 0.092

ables are most significant to the predictive performance, we look at the last two rows of the table. Here we consider the fact that the tacrolimus variable that most adds to the static set performance is in both cases an expression of tacrolimus concentration. This tells us that the tacrolimus concentration values contribute the most to the prediction of rejection probability at t = 168h when compared to other tacrolimus measurements.

Supplementary Results - IPW experiments

Initially the denominator predictor, a comprehensive history encompassing the target history as well as all variables influencing tacrolimus values was defined, acknowledging the uncertainties in the causal graph (adding more variables related to tacrolimus instead of just the obvious confounders helps to account for the uncertainty in the causal graph). This predictor lead to overconfident probability distribution and, more importantly, a high weight mean (the ideal mean of standardized weights is 1). Regularization was employed to avoid this, but the amount of regularization needed to generate weights with a mean of 1 lead us to believe there were better solutions.

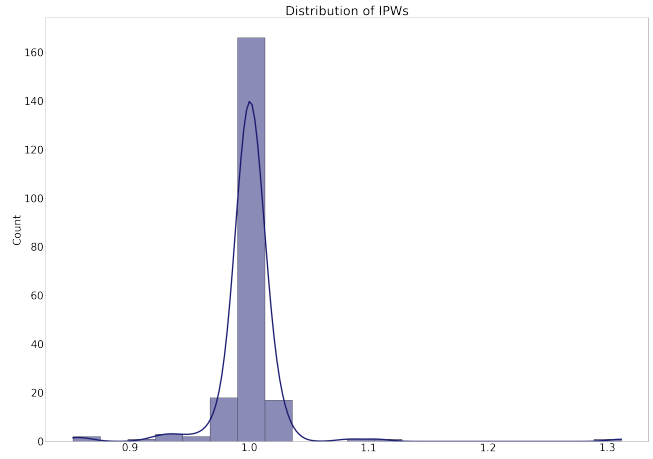
Since these results are indicative of incorrect model specification or lack of positivity [Cole and Hernán, 2008], we decided to investigate. A smaller set was chosen that only considers the confirmed confounders: age and the use of steroid treatment. Since there is no information on steroid treatment, the variable `co_immuno` that indicates the use of other immunosuppressive treatment was used as the only that could have some information on the use of steroids. This new set achieved much more stable results, as evidenced by Table 4 and Figure 8.

Supplementary Results - MSM

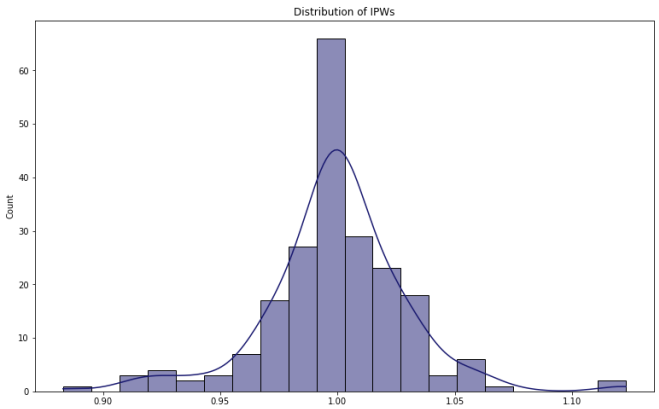
Here an IPW weighted Logistic Regression was trained on the following coefficients: the treatment ($\%HOT$), a demographic factor deemed related to the outcome (Age) and the relationship between the two ($\%HOTxAge$). This procedure was bootstrapped over 500 samples. The objective was to understand if different demographic groups manifest different causal relationships between treatment and outcome. To do this we analyse the interaction between age and $\%HOT$, as shown in Figures 10 and 11. Here we can see that the coefficient of the interaction has a relatively small absolute value and a considerable amount of uncertainty around it, which leads to the relative effect is not significant (the probability of rejection does not change for different combinations of treatment and age). Future research might separate the patients by different demographic groups, or further investigate age by separating patients into groups.

Table 4: IPW regularization and confounder set experiment results. Here C is the inverse of regularization strength, and preset is 1. We can see that the confounder set exhibits much more stable results, as the means are all around 1. For the whole set, the regularization strength needs to reach very high values to get to more stable weight results, leading to the thought that this model is misspecified.

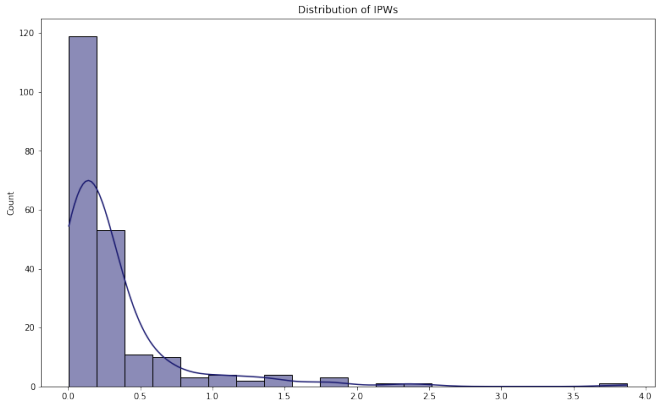
	whole set				confounders set			
	C = 1000	preset	C = 0.001	C = 0.0001	C = 1000	preset	C = 0.001	C = 0.0001
mean	0.3141	0.2881	0.5069	0.7466	1.0000	1.0000	0.9995	0.9995
std	0.4681	0.4592	0.3287	0.1798	0.0306	0.0283	0.0305	0.0055
min	0.0044	0.00467	0.0266	0.2449	0.8523	0.8643	0.8829	0.9788
max	4.0766	0.0051	3.0114	1.0946	1.3117	1.2806	1.1229	1.0214



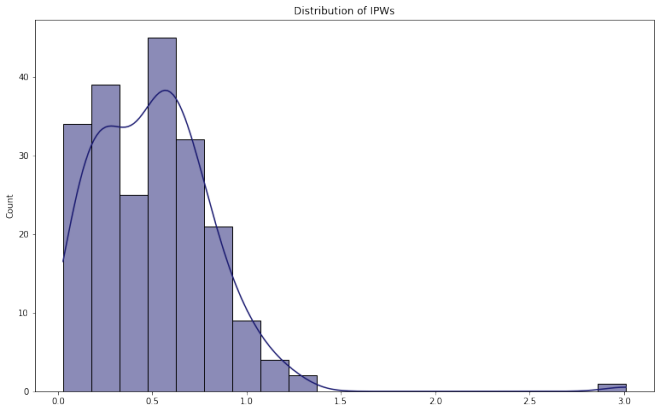
(a) IPW histogram for counfounder set, C = 1000



(b) IPW histogram for counfounder set, C = 0.001

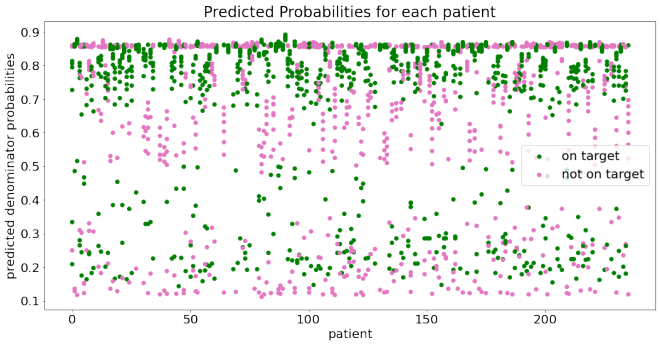


(c) IPW histogram for whole set, C = 1000

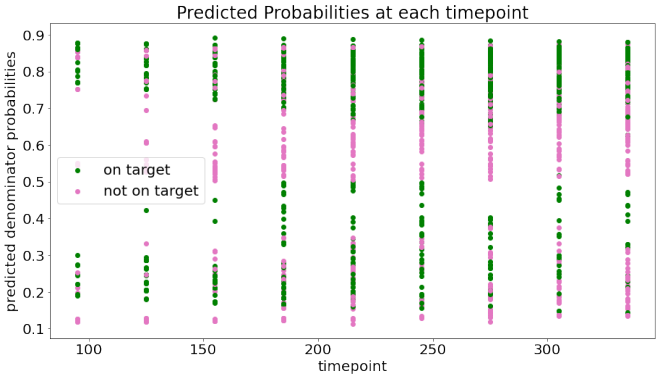


(d) IPW histogram for whole set, C = 0.001

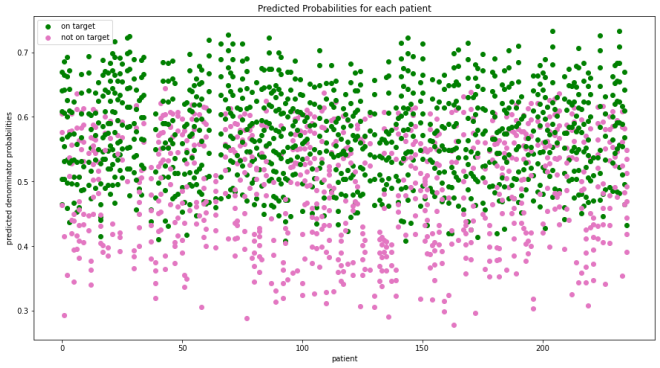
Figure 8: Weight distributions for different IPW configurations. As we can see the confounder set has results centered around 1, which was the objective.



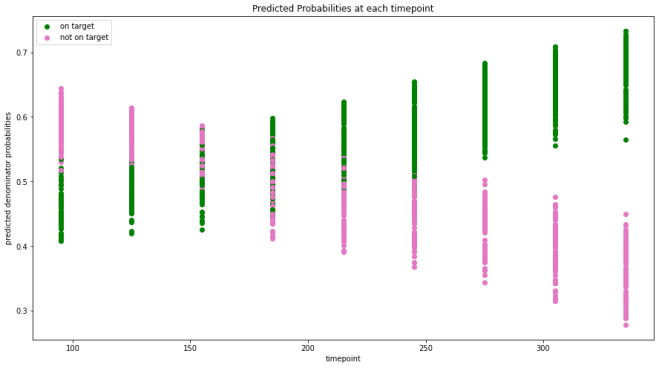
(a) Denominator distribution per patient for counfounder set, C = 1000



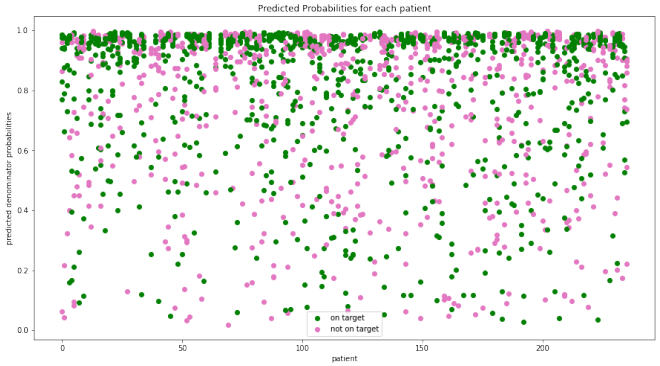
(b) Denominator distribution per timepoint for counfounder set, C = 1000



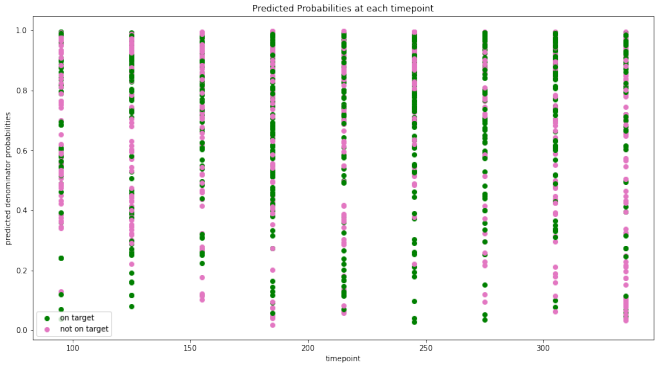
(c) Denominator distribution per patient for counfounder set, C = 0.0001



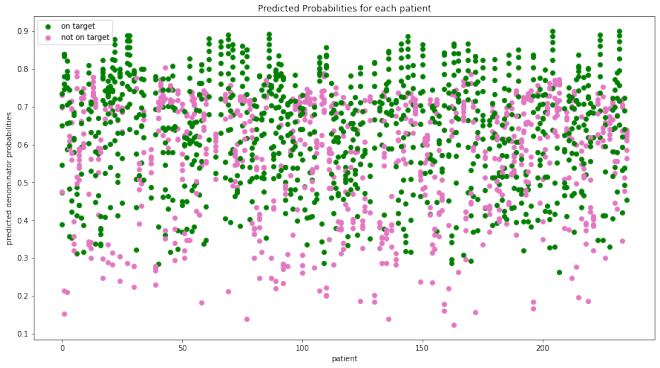
(d) Denominator distribution per timepoint for counfounder set, C = 0.0001



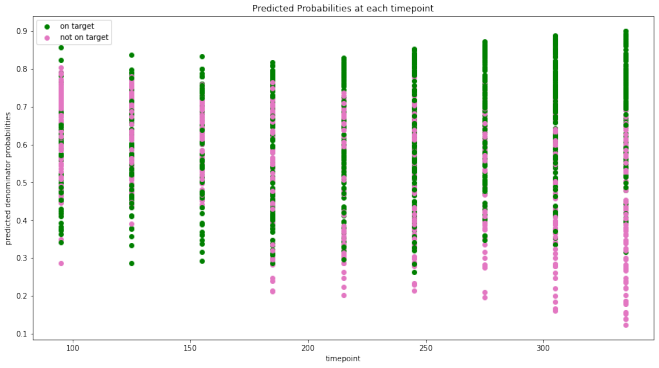
(e) Denominator distribution per patient for whole set, C = 1000



(f) Denominator distribution per timepoint for whole set, C = 1000



(g) Denominator distribution per patient for whole set, C = 0.001



(h) Denominator distribution per timepoint for whole set, C = 0.001

Figure 9: Denominator distributions for different IPW configurations. The objective here is to obtain results tending towards 1 while avoiding overfitting, which would look something like a horizontal line around 1

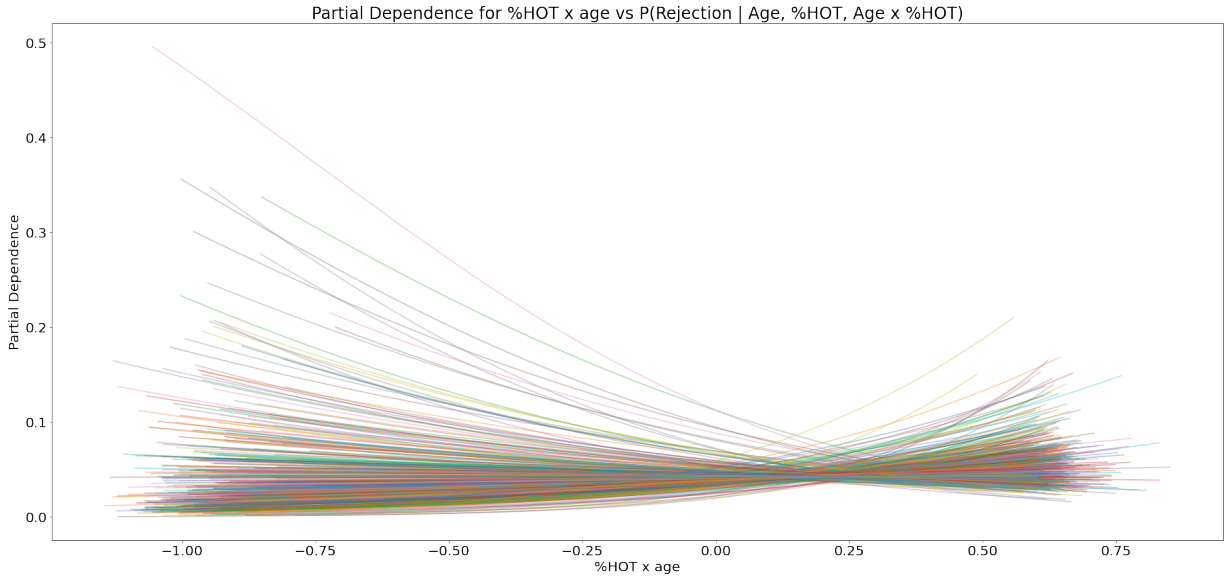


Figure 10: Partial Dependence plots for Age x %HOT. We can see that the curve is generally flat, except in the outliers where it tends to decline.

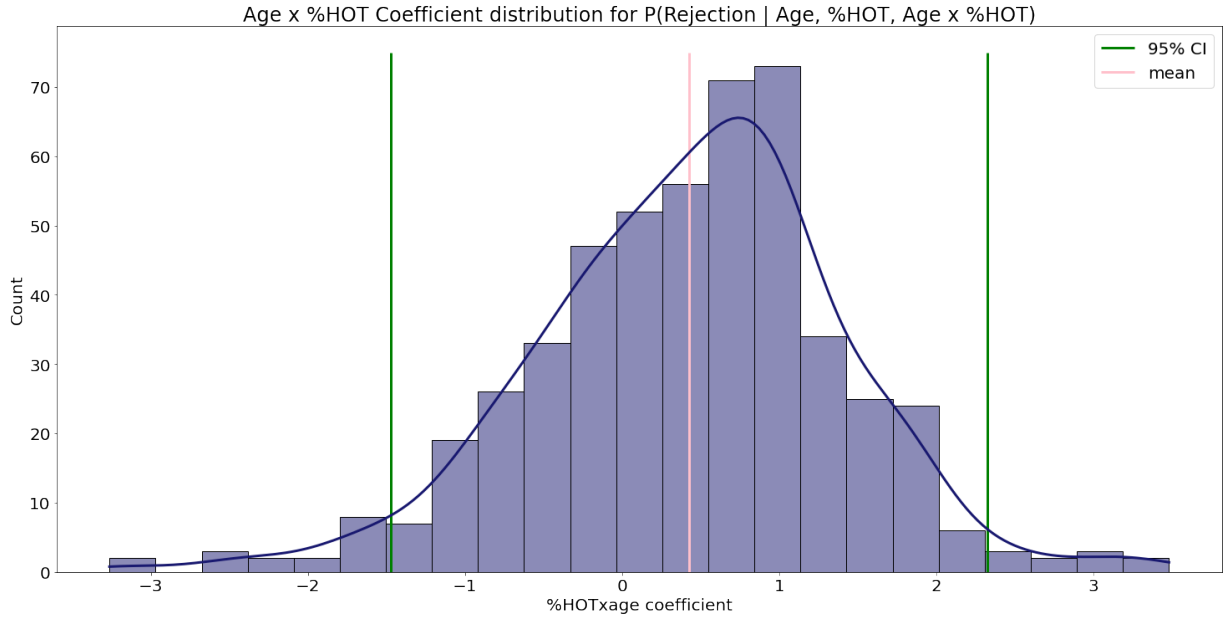


Figure 11: Age x %HOT coefficients for P(Rejection—Age, %HOT, Age x %HOT) model. The average has a small absolute values, meaning FINISH THIS