Modeling Functional Relationships in Causal Graphs and Estimating Individualized Interventions: Neural Causal Models (TRAM-DAGs) and Conditional Average Treatment Effects

Master Thesis in Biostatistics (STA495)

by

Mike Krähenbühl

Matriculation number: 18-652-149

supervised by

Prof. Dr. Beate Sick

Prof. Dr. Oliver Dürr, HTWG Konstanz

Zurich, July 2025

Modeling Functional Relationships in Causal Graphs and Estimating Individualized Interventions:

Neural Causal Models (TRAM-DAGs) and Conditional Average Treatment Effects

Mike Krähenbühl

Version July 3, 2025

Contents

Preface

In the introduction part, my aim is to give a summary of important concepts of causal inference and causal models. Further, I motivate the importance for methods that allow to draw causal conclusions from observational data in contrast to randomized controlled trials (RCTs) and that the proposed framework of tram-dags (cite sick duerr) can be used as such a tool.

In the methods section, I give a detailed description of the tram-dag framework and how it works by illustrating it on a simple simulated example. I also show for what kind of causal queries the model can be used for. Although it is not a topic for observations data, I discuss how the model can be used for estimating the individualized treatment effect (ITE).

In the results section, I will first show results from simulation studies where the ground truth is known. Second, I will show results on a real-world example in the setting of climate data. Third, I present the results of the ITE estimation.

In the final section, I will discuss the results and give a conclusion about the advantages and limitation of this framework and provide an outlook.

Mike Krähenbühl
July 2025

Abstract

Abstract here

Keywords: Neural Causal Models (TRAM-DAGs), Structural Causal Models, Counterfactuals, Transfrmation Models, Observational Data, Heterogeneous Treatment Effects, Individualized Treatment Effects (ITE), Conditional Average Treatment Effects (CATE)

Chapter 1

Introduction

1.1 Motivation

The most important questions in research are mostly not associational, but causal (Pearl, 2009a). They concern the effects of interventions — such as the impact of a treatment — or seek explanations for observed outcomes, such as identifying which disease caused certain symptoms. They also include hypothetical scenarios; for example: what would the GDP have been if interest rates had increased by 25 instead of 75 basis points? Answering such questions requires causal reasoning and demands an understanding of the underlying data-generating process. Purely associational approaches are typically not sufficient to draw valid causal conclusions.

The gold standard for estimating the causal effect of an intervention on an outcome is the randomized controlled trial (RCT) (Hariton and Locascio, 2018). In this prospective study design, participants are randomly assigned to either the treatment or control group. Randomization aims to eliminate the influence of potential confounding variables, ensuring that treatment groups are balanced with respect to baseline characteristics. This allows for an unbiased estimation of the causal effect. Despite their strengths, RCTs have several limitations. They are often expensive and time-consuming to plan and execute. Moreover, the results may not generalize well to the population of interest, as individuals who volunteer or are accepted for trials are not always representative of the target group. Additionally, RCTs typically estimate an average treatment effect (ATE) on a sample, which is the difference in mean outcomes between treatment arms (Nichols, 2007). However, individual patients may respond differently to the treatment, depending on their unique characteristics. In the context of personalized medicine, it is therefore crucial to have an estimate of treatment effects at the individual level. Another central limitation of RCTs is that in many scenarios they can simply not be conducted due to ethical or practical reasons. For example, an RCT is only ethical in the case of clinical equipoise, which means that there is uncertainty about the (superiority) of one of the two treatment arms (Freedman, 1987). It is not acceptable to treat one group with the assumed inferior treatment. The same is true for obviously harmful interventions, like smoking or drinking alcohol. In these cases, it is not possible to conduct an RCT to estimate the causal effect of smoking on lung cancer.

For these reasons, much of research aims to make causal inference from observational data, using non-experimental or quasi-experimental designs. Unlike RCTs, these settings do not involve randomization to treatment, which introduces challenges due to confounding. Methods for causal inference from observational data aim to correctly control for such confounders to enable valid causal conclusions. Recently, Sick and Dürr (2025) proposed the TRAM-DAGs framework, which estimates the functional form of causal relationships in a known causal graph based on observational or RCT data and make subsequent causal queries. In this thesis we further analyze and apply this method.

As mentioned earlier, an application where causal inference is of particular importance is the estimation of personalized treatment effects. In personalized medicine, this is referred to as the individualized treatment effect (ITE) or conditional average treatment effect (CATE), while in business and marketing contexts, the term uplift modeling is often used (Gutierrez and Gérardy, 2017; Zhao and Harinen, 2020). These concepts refer to the difference in potential outcomes under different treatments, assessed at the level of individuals or subgroups. Such estimates are critical in settings where treatment responses vary significantly between individuals. For clinical decision-making, tailoring therapies to individual characteristics can lead to more effective and efficient care. The importance of estimating individual-level effects is not limited to medicine. It also is of high interest in marketing, where campaigns can be precisely targeted to maximize impact and minimize adverse responses. Consider, for instance, the decision of whether to send a push notification (treatment) to a customer. Some customers might be persuadables, who will respond positively only if treated. Others, in contrast, might have responded positively without the intervention but are negatively affected by it – for example, a customer who is reminded of a forgotten subscription and, as a result, decides to cancel it. In this context, identifying persuadables is valuable, while treating the latter may be counterproductive. This illustrates the need to understand treatment effects at a granular level to guide individualized decisions. Various methods have been proposed to estimate individualized treatment effects, yet this task remains challenging. The fundamental problem is that only one of the two potential outcomes can ever be observed for any given individual, making the estimation of treatment effects inherently more difficult than standard predictive modeling.

1.2 Key concepts of causal inference

Causal relationships can be represented by a directed acyclic graph (DAG), as shown in Figure 1.1(a). The variables, or nodes, are connected by directed edges, which represent causal dependencies.

Questions in causal inference are typically classified into one of the three levels of Pearl's hierarchy of causation (Pearl, 2009b). Level 1 corresponds to observational queries, expressed as conditional probabilities $P(Y \mid X)$, which can be answered directly from the joint distribution $P(Y \cap X)$. Level 2 involves interventional queries, such as $P(Y \mid do(X))$, which describe the probability when actively setting a variable X to a particular value. Unlike observational queries, answering interventional questions requires knowledge of the underlying causal structure. Level 3 addresses counterfactual reasoning, which poses the greatest challenge. These are hypothetical what-if questions that require reasoning about outcomes under alternative realities. For example, if a patient received a treatment and died, the factual outcome is death under the received treatment. The counterfactual would be the outcome that would have occurred had the patient received a different treatment.

Some statisticians argue that counterfactuals — being unobservable and untestable — are of limited scientific value and may be regarded as metaphysical (Dawid, 2000). Nevertheless, there are important practical questions that require the analysis of such counterfactuals.

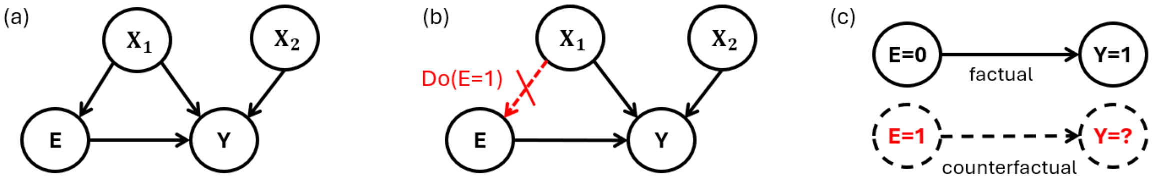


Figure 1.1: Example for the three levels of Pearl's hierarchy of causation. (a) DAG for observational data. (b) DAG when making a do-intervention by fixing the variable E at a certain value. c) Observed factual outcome and corresponding counterfactual query.

To illustrate Pearl's three levels of causality, I consider a simplified example involving the exposure Exercise (E), the outcome Heart Disease (Y), the confounder Age (X_1) and the additional covariate Smoking (X_2). I assume that exercise reduces the risk of heart disease, but both variables are also influenced by age. Figure 1.1(a)–(c) illustrates the corresponding scenarios.

Level 1: Observational ("seeing"): We observe the joint distribution of variables without intervention. Example: What is the probability of heart disease given that a person exercises?

$$P(Y = 1 \mid E = 1)$$

This can be estimated directly from data by conditioning on $E = 1$ and computing the frequency of Y. However, such an estimate does not account for confounding variables like age.

Level 2: Interventional ("doing"): We consider the effect of actively intervening in the system. Example: What is the probability of heart disease if everyone were made to exercise, regardless of age or smoking status?

$$P(Y = 1 \mid \mathrm{do}(E = 1))$$

Answering this requires assumptions about the underlying causal structure.

Level 3: Counterfactual ("imagining"): We ask what would have happened under different circumstances – that is, we imagine an alternative scenario for the same individual. Example: For a person who does not exercise and has heart disease, would they still have had heart disease if they had exercised?

$$P(Y_{E=1} \mid E = 0, Y = 1)$$

Here, $Y_{E=1}$ represents the counterfactual outcome under positive exposure. Counterfactual queries require a structural causal model (SCM) and cannot be answered from data alone.

Here, $Y_{E=1}$ represents the counterfactual outcome under positive exposure. Counterfactual queries cannot be answered from observational data alone; they require a structural framework that explicitly models the data-generating process.

For this, the concept of DAGs can be extended to structural causal model (SCM). A set of structural equations of the form $X_i = f_i(\mathrm{pa}(X_i), Z_i), \quad i = 1, \ldots, n$ build a structural causal model (Pearl, 2009b). $\mathrm{pa}(X_i)$ denotes the direct causal parents of X_i, and Z_i is an exogenous noise variable. These exogenous variables capture latent factors that influence X_i but are not explicitly modeled. By convention, the Z_i are assumed to be mutually independent.

Each function f_i — which may be nonlinear — defines how the value of X_i is generated from its parents and the corresponding noise term. A source node X_j without any parents is modeled as $X_j = f_j(Z_j)$. Once all structural equations and noise variables are specified, the model is fully deterministic in the sense that each variable is a fixed function of its parents and its own exogenous noise. The randomness in the system arises entirely from these independent noise terms, which encode unobserved factors. This functional representation makes it possible to compute interventional distributions and evaluate counterfactual outcomes. These aspects are discussed in detail in Section 2.1.4.

In this thesis, we do not focus on discovering the underlying causal graph. Such a structure may be obtained through structure learning algorithms or determined from expert knowledge. Instead, we assume the graph is known and concentrate on estimating the functional form of the relationships between variables – that is, the structural equations that define the SCM.

Various approaches exist for estimating the functions f_i that constitute an SCM, depending on the assumptions made about the data and the model class. These methods are discussed in the next section.

A simple approach to modeling the structural equations is linear regression, which assumes Gaussian error terms Z_i and linear functional forms f_i. Classical statistical methods of this kind are typically well-defined, computationally efficient, and offer interpretable parameters.

However, they rely on strong assumptions about the underlying data-generating mechanism — such as linearity and homoscedasticity — which may not hold in practice. Violations of these assumptions can lead to biased or misleading results.

Alternatively, more flexible approaches based on neural networks have gained popularity for estimating structural equations. These models are capable of approximating complex, nonlinear relationships and capturing complicated interactions between variables with minimal bias. Their flexibility, however, often comes at the cost of reduced interpretability and, in some cases, limited applicability to non-continuous or mixed data types. Poinsot *et al.* (2024) provided an overview of deep structural causal models and their use in counterfactual inference.

The TRAM-DAG framework proposed by Sick and Dürr (2025) builds a bridge between these classical and neural-network-based modeling approaches, by combining interpretable transformation models with the flexibility of neural networks. At its core, the structural equations are modeled using transformation models (Hothorn *et al.*, 2014), a flexible class of distributional regression methods. These models were subsequently extended to deep transformation models (Deep TRAMs) by Sick *et al.* (2021), enabling the use of neural networks to parameterize conditional distributions in a customizable way. In the TRAM-DAG framework, these deep TRAMs are applied according to a known causal graph, allowing the model to be fitted to observational data and used to answer causal queries across all three levels of Pearl's hierarchy. The framework is introduced in more detail in Section 2.1.3.

1.3 Goals and contributions

This thesis contributes to the further exploration of the TRAM-DAG framework and to adressing challenges in the estimation of personalized treatment effects.

The first part of this thesis focuses on a systematic analysis and extension of TRAM-DAGs. This includes applying the model across a variety of settings, such as different data types, model complexities, and neural network configurations (e.g., activation functions, batch normalization, dropout). Most analyses are conducted on simulated data to know the underlying data-generating process, but the model is also applied to real-world data to demonstrate its practical utility.

The second focus of the thesis is on the estimation of personalized treatment effects. Recent work by Chen *et al.* (2025) showed that most causal machine learning models trained on RCT data failed to generalize when evaluated out of sample. In this thesis, we replicate some of their work by applying various models, including TRAM-DAGs, to the same data and analyzing whether we come to a similar conclusion. We further investigate why individualized treatment effect (ITE) estimation can fail in such settings, and under which conditions reliable estimates can be obtained. In addition, we demonstrate that TRAM-DAGs can be effectively used to estimate ITEs even in non-randomized observational settings, provided that the causal graph is known and fully observed. In doing so, we explore the potential of TRAM-DAGs as a framework for answering complex causal questions across different levels of Pearl's hierarchy.

Formally, we aim to answer following research questions in this thesis:

- How can TRAM-DAGs be applied under different scenarios such as ordinal predictors, scaled vs. raw variables or allowing for interactions between variables.

- Do we obtain similar results when estimating ITEs on a real-world RCT dataset, as reported by Chen *et al.* (2025)?

- What are possible reasons for the failure of ITE estimation in some cases when causal machine learning models are validated out of sample?

- How can TRAM-DAGs be used to estimate ITEs in both randomized controlled trials and observational settings involving confounding and mediating variables?

With this work, we aim to contribute to the important and evolving field of causal inference in observational settings and to the challenging task of estimating individualized treatment effects.

Chapter 2

Methods

This chapter introduces the methodological foundations and experimental designs used in this thesis. Section 2.1 presents the concept and functionality of TRAM-DAGs, along with the necessary theoretical background. Section 2.2 provides the framework for estimating individualized treatment effects. Finally, Sections 2.3–2.6 describe the four experiments conducted to address the research questions outlined in Section 1.3.

2.1 TRAM-DAGs

The goal of TRAM-DAGs is to estimate the structural equations according to the causal order in a given DAG in a flexible and possibly still interpretable way in order to sample observational and interventional distributions and to make counterfactual statements. The estimation requires data and a DAG that describes the causal structure. It must be assumed that there are no hidden confounders. TRAM-DAGs estimate for each variable X_i a transformation function $Z_i = h_i(X_i \mid pa(X_i))$, where Z_i is the noise value and $pa(X_i)$ are the causal parents of X_i. The important part here is that we can rearrange this equation to $X_i = h_i^{-1}(Z_i \mid pa(x_i))$ to get to the structural equation. The transformation functions h are monotonically increasing functions that are a representation of the conditional distribution of X_i on a latent scale. They are based on the idea of transformation models as introduced by Hothorn *et al.* (2014) but were extended to deep trams by Sick *et al.* (2021). In the following sections I review the most important ideas of these methods as they are the essential components of TRAM-DAGs.

2.1.1 Transformation Models

Transformation models are a flexible distributional regression method for various data types. They can be for example specified as ordinary linear regression, logistic regression or proportional odds logistic regression. But Transformation models further allow to model conditional outcome distributions that do not even need to belong to a known distribution family of distributions by model it in parts flexibly. This reduces the strength of the assumptions that have to be made.

The basic form of transformation models can be described by

$$F(y|\mathbf{x}) = F_Z(h(y \mid \mathbf{x})) = F_Z(h_I(y) - \mathbf{x}^\top \boldsymbol{\beta}) \tag{2.1}$$

, where $F(y|\mathbf{x})$ is the conditional cumulative distribution function of the outcome variable Y given the predictors $\mathbf{x}$. $h(y \mid \mathbf{x})$ is a transformation function that maps the outcome variable y onto the latent scale of Z. F_Z is the cumulative distribution function of a latent variable Z, the so-called inverse-link function that maps $h(y \mid \mathbf{x})$ to probabilities. In this basic version, the transformation function can be split into an intercept part $h_I(y)$ and a linear shift part $\mathbf{x}^\top\boldsymbol{\beta}$, where the vector $\mathbf{x}$ are the predictors and $\boldsymbol{\beta}$ are the corresponding coefficients.

If the latent distribution Z is chosen to be the standard logistic distribution, then the coefficient β_i can be interpreted as log-odds ratios when increasing the predictor x_i by one unit, holding all other predictors unchanged. This means that an increase of one unit in the predictor x_i leads to an increase of the log-odds of the outcome Y by $\boldsymbol{\beta}$. The additive shift of the transformation function means a linear shift on the latent scale (herer log-odds). The following transformation to probabilities by F_Z potentially leads to a non-linear change in the conditional outcome distribution on the original scale. This means not only is the distribution shifted, also its shape can change to some degree based on the covariates. More details about the choice of the latent distribution and the interpretation of the coefficients are provided in the appendix XXX.

For a continuous outcome Y the intercept h_I is represented by a bernstein polynomial, which is a flexible and monotonically increasing function

$$h_I(y) = \frac{1}{M+1} \sum_{k=0}^{M} \vartheta_k \, \mathrm{B}_{k,M}(y) \tag{2.2}$$

, where ϑ_k are the coefficients of the bernstein polynomial and $\mathrm{B}_{k,M}(y)$ are the Bernstein basis polynomials. More details about the technical implementation of the bernstein polynomial in the context of TRAM-DAGs is given in the appendix XXX.

For a discrete outcome Y the intercept h_I is represented by cut-points, which are the thresholds that separate the different levels of the outcome. For example, for a binary outcome Y there is one cut-point and for an ordinal outcome with K levels there are $K-1$ cut-points. The transformation model is given by

$$P(Y \leq y_k \mid \mathbf{X} = \mathbf{x}) = F_Z(\vartheta_k + \mathbf{x}^\top \boldsymbol{\beta}), \quad k = 1, 2, \ldots, K-1 \tag{2.3}$$

A visual representation for a continuous and discrete (ordinal) outcome is provided in Figure 2.1.

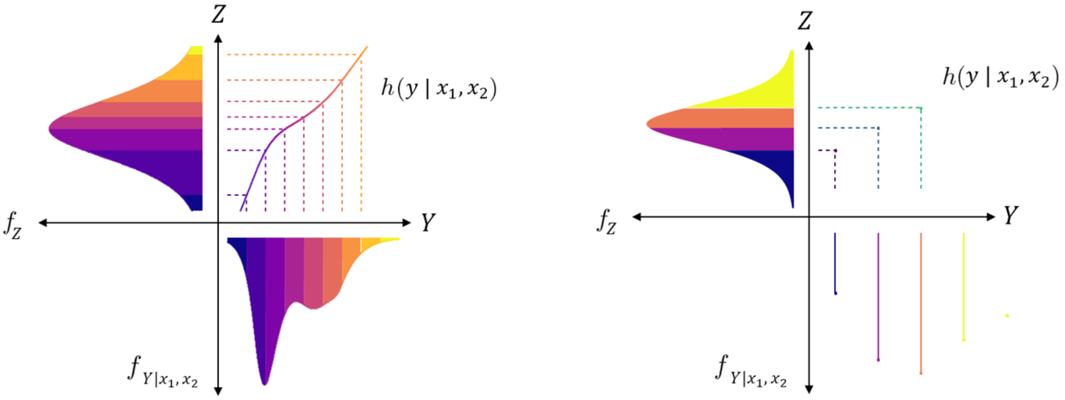


Figure 2.1: Left: Example of a transformation model for a continuous outcome Y with a smooth transformation function. **Right:** Example of a transformation model for an ordinal outcome Y with 5 levels. The transformation function consists of cut-points that separate the probabilities for the levels of the outcome. In both cases the latent distribution Z is the standard logistic and the predictors $\mathbf{x}$ induce a linear (vertical) shift of the transformation function.

To estimate the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}$ the negative log likelihood (NLL) is minimized. The NLL is defined as

$$\mathrm{NLL} = -\frac{1}{n} \sum_{i=1}^{n} l_i(\boldsymbol{\beta}, \boldsymbol{\vartheta}) = -\frac{1}{n} \sum_{i=1}^{n} \log(f_{Y|\mathbf{X}=\mathbf{x}}(y_i)) \tag{2.4}$$

where $l_i(\boldsymbol{\beta}, \boldsymbol{\vartheta})$ is the log-likelihood of the i-th observation, $l_i(\boldsymbol{\beta}, \boldsymbol{\vartheta}) = f_{Y|\mathbf{X}=\mathbf{x}}(y_i)$ is the conditional density function of the outcome variable Y given the predictors $\mathbf{x}$ under the current parameterization. I provide the full derivation in the appendix xxx.

For the remainder of this thesis, I rely on the idea of these transformation models to model the conditional distribution functions represented by the transformation functions of the respective variables. The standard logistic distribution is used as F_Z, which results in a logistic transformation model.

2.1.2 Deep TRAMs

The transformation models as discussed before were extended to deep TRAMs using a modular neural network (Sick *et al.*, 2021). The goal is to get a parametrized transformation function of the form **??**.Each part, the intercept $h_I(X_i)$, the linear shift $\mathbf{x}_L^\top \boldsymbol{\beta}_L$ and the complex shift $f_C(\mathbf{x}_C)$ are assembled by the outputs of the individual neural networks. The user can specify the level of complexity the parents $pa(X_i)$ have on the transformaiton funciton. Figure 2.2 illustrates the case for a SI-LS-CS model.

$$h(y \mid \mathbf{x}_L, \mathbf{x}_C) = h_I(y) + \mathbf{x}_L^\top \boldsymbol{\beta}_L + f_C(\mathbf{x}_C) \tag{2.5}$$

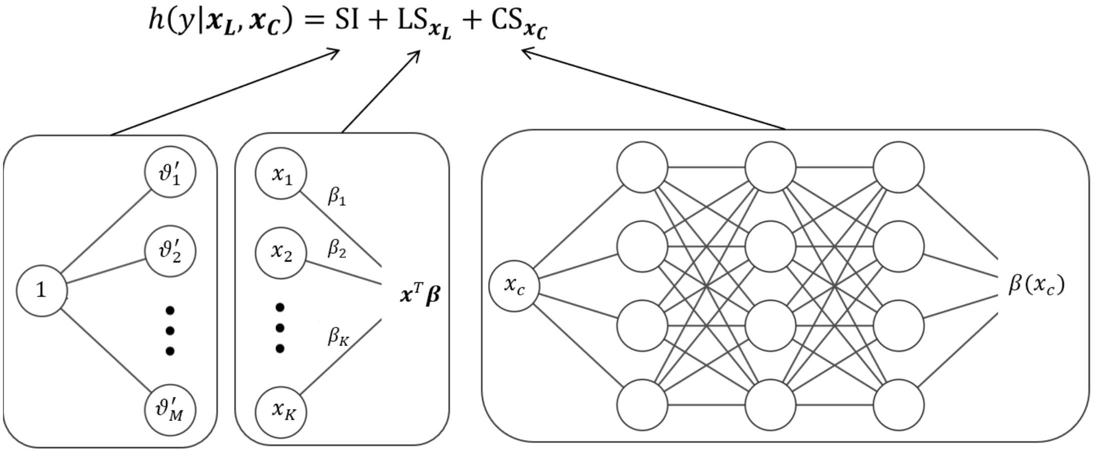

$h(y|\boldsymbol{x_L}, \boldsymbol{x_C}) = \mathrm{SI} + \mathrm{LS}_{\boldsymbol{x_L}} + \mathrm{CS}_{\boldsymbol{x_C}}$
1
ϑ'_1
ϑ'_2
ϑ'_M
x_1
x_2
x_K
β_1
β_2
β_K
$\boldsymbol{x}^T\boldsymbol{\beta}$
x_c
$\beta(x_c)$

Figure 2.2: Modular deep transformation model. The transformation function $h(y \mid \mathbf{x})$ is constructed by the outputs of three neural networks.

Intercept the shape of the transformation function at the baseline configuration $\mathbf{x}_L^\top \boldsymbol{\beta}_L = 0$ and $f_C(\mathbf{x}_C) = 0$ is determined by the intercept $h_I(y)$. For a continuous outcome the intercept is represented by a smooth bernstein polynomial and in the discrete case by cut-points. In either case the parameters ϑ are obtained as output nodes of the neural network. A simple intercept (SI) is the case where the parameters ϑ do not depend on the any explanatory variables. The neural network thereby only takes a constant as input and directly outputs the parameters ϑ. To make the intercept more flexible, the intercept can also depend on the explanatory variables. In this case the complex intercept (CI) models the intercept $\vartheta(x)$ by taking the predictors x as input to a neural network with some hidden layers. This allows the intercept to change with the value of the predictors. Depending on the assumptions, predictors can be used in the complex intercept, or only a subset of them. A detailed explanation of the construction of the bernstein polynomial is given in appendix XXX.

Linear shift If the predictors should have a linear effect on the transformation function, it can be modelled by a linear shift (LS). For this part the neural network without hidden layers and without biases takes the linear predictors $pa(X_i)$ as input and generates a single output node with a linear activation function. This results in the linear combination $\mathbf{x}_L^\top \boldsymbol{\beta}_L$ and it induces a linear vertical shift of the transformation function. The weights $\boldsymbol{\beta}_L$ are the interpretable coefficients of the linear shift. For the logistic transformation model, they are interpreted as log-odds-ratios. The interpretation is further described in the appendix 6.2.

Complex shift If the transformation function should be allowed to be shifted vertically in a non-linear manner, a complex shift (CS) can be applied. The predictor variables are inputted in a (deep) neural network with at least one hidden layer and non-linear activation functions such as sigmoid or ReLU. A single output node with $f_C(X_C)$ is obtained. With a complex shift, also interactions between predictor variables can be captured by giving the interacting variables into the same neural network.

Level of complexity One practical feature of these modular deep TRAMs is that one can specify, which predictors should have a linear or complex shift effect on the transformation function or that predictors are even allowed to deterimine the shape of the transformation function by a complex intercept. Herzog *et al.* (2023) predicted the ordinal functional outcome three months after stroke by using semi-structured data that included tabular predictors and images. The two data modalities can be included in a single deep TRAM by modeling the part of the images with a CNN.

The estimated distribution function is invariant with respect to the choice of the inverse-link function F_Z (scale of latent distribution) in an unconditional (Hothorn *et al.*, 2018) or fully flexible (CI) setting. However, as soon as restrictions are placed on the influence of the predictors (LS, CS), this leads to assumptions about the scale of the dependency. Which latent distribution should be chosen depends on following factors: (i) the intended complexity of the model, (ii) the assumptions about the data generating process, (iii) the conventional, widely used, scale of interpretation for the specific problem. If the coefficients β in the linear shift term should be interpreted as log odds ratios, then the standard logistic distribution is appropriate. For log hazard ratios it would be the minimum extreme value distribution. There exist plenty of other alternatives.

(The optimal scale could be found by comparing the likelihoods of the model under different latent distributions.)

Parameter estimation The parameters of the neural networks are learned by minimizing the negative log-likelihood (NLL) of the conditional deep TRAM. The learning process is started with a random parameter configuration and the outputs of the neural networks are used to assemble the NLL of the transformation model. The NLL is then iteratively minimized by adjusting the parameters by the Adam optimizer (Kingma and Ba, 2015) until they eventually converge to the optimum state. Additionally, methods to prevent overfitting — such as dropout, early stopping, or batch normalization — can be applied. These techniques are particularly important in more complex networks to ensure that the model generalizes well to out-of-sample data. In the hidden layers, non-linear activation functions such as ReLU or sigmoid are applied.

2.1.3 TRAM-DAGs: Deep TRAMS applied in a causal setting

In TRAM-DAGs these deep transformation models are applied in a causal setting. We assume a pre-specified DAG which defines the causal dependence. Then we estimate the distribution of each node by a transformation model that is conditional on its parents. Figrue 2.3 illustrates the basic idea of a TRAM-DAG where a DAG with 3 variables, without hidden confounder, is assumed to be known. The arrows in the DAG indicate the causal dependencies between the variables. The transformation models are constructed by a modular neural network. The assumed influence from the parent variables has to be specified as SI, LS or CS. In this example, X_1 is a continuous source node that acts as parent of X_2 and X_3. For a source node the transformation function only consists of a simple intercept (SI). X_2 is also continuous and its transformation function can be shifted additively (LS) by the value of X_1. X_3 is an ordinal variable with 4 levels and its transformation function depends on the values of X_1 (LS) and X_2 (CS). The cut-points $h(x_3 \mid x_1, x_2)$ represent the cumulative probabilities on the log-odds scale of the first 3 levels of X_3, where the probability of the last level $K = 4$ is the complement of the previous levels k_{1-3}.

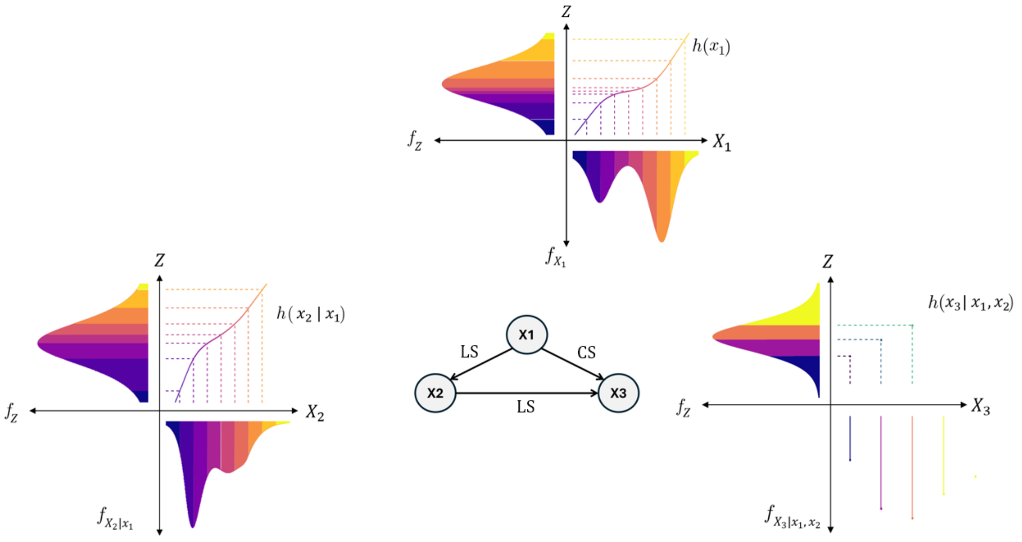


Figure 2.3: Example of a TRAM-DAG with three variables X_1, X_2 and X_3. The transformation functions are represented by the modular neural networks. The arrows indicate the causal dependencies between the variables.

This DAG with the assumed dependencies can be described by an adjacency matrix 2.6, where the rows indicate the source and the columns the target of the effect:

$$\mathbf{MA} = \begin{bmatrix} 0 & \mathrm{LS} & \mathrm{LS} \\ 0 & 0 & \mathrm{CS} \\ 0 & 0 & 0 \end{bmatrix} \tag{2.6}$$

To apply the framework of TRAM-DAGs on this example, we assume to have observational data that follows the structure of the adjacency matrix 2.6. In practice, the DAG is either defined by expert knowledge or by some sort of structure finding algorithm (XXX cite methods). Then we want to estimate the conditional distribution function of each variable by a deep TRAM so that we can sample from the distributions and make causal queries. The conditional distribution functions are given by

$$\begin{aligned} X_1 &\sim F_Z(h_I(x_1)) \\ X_2 &\sim F_Z(h_I(x_2) + \mathrm{LS}_{x_1}) \\ X_3 &\sim F_Z(h_I(x_3) + \mathrm{LS}_{x_1} + \mathrm{CS}_{x_2}) \end{aligned}$$

Construct Modular Neural network

As discussed in the section 2.1.2, the transformation functions are constructed by a modular neural network. The inputs are the variables in the system as well as the adjacency matrix 2.6 which controls the information flow and assures that only valid connections according to the causal dependence are made. Discrete variables with few categories are dummy encoded, and continuous variables should be scaled before feeding them in the neural network. The encoding and the effect of scaling on the interpretation of parameters is discussed in the appendix (6.3 and 6.4). Scaling the input variables, meaning to bring the variables onto a zero-mean and one-variance, can remove the pattern in marginal variance which some structure learning algorithms rely on (Reisach *et al.*, 2021). However, since our analysis does not require to find the structure and already assumes a known DAG, this is not a problem. Once the input variables are prepared and the structure is defined by the adjacency matrix, the architecture of the neural network for the complex shift and complex intercept has to be specified. These are factors such as depth, width, activation function, and whether dropout or batch normalization should be used. These considerations depend on the assumed complexity of the shifts. The outputs of the neural networks are the three components for the transformation function (SI, LS, CS) for each variable. These components are assembled to the transformation functions. Finally, the loss is defined as the negative log likelihood, which the model aims to optimize to estimate the optimal parameterization. The estimated parameters ***beta*** in the linear shifts are interpretable as log-odds-ratios when changing the value of the respective parent by one unit, leaving all others unchanged.

2.1.4 Sampling from TRAM-DAGs

Observational sampling Once the TRAM-DAG is fitted on data, it can be used to sample from the observational or interventional distribution or to make counterfactual queries. The structural equations $X_i = f(Z_i, \mathrm{pa}(X_i))$ are represented by the inverse of the conditional transformation functions $h^{-1}(Z_i \mid \mathrm{pa}(X_i))$ because $Z_i = h(X_i \mid \mathrm{pa}(X_i))$. The sampling process from the observational distribution for one iteration (one observation of all variables in the DAG) is described in the pseudocode 1 and illustrated in Figure 2.4. The process is repeated for the desired number of samples.

Algorithm 1 Generate a samples from the TRAM-DAG

```
Given: A fitted TRAM-DAG with structural equations $X_i = f(Z_i, \mathrm{pa}(X_i))$, where $Z_i = h(X_i \mid \mathrm{pa}(X_i))$
for each node $X_i$ in topological order do
    Sample latent value $z_i \sim F_{Z_i}$                              ▷ e.g., rlogis() in R
    if $X_i$ is continuous then
        Compute $x_i = h^{-1}(z_i \mid \mathrm{pa}(x_i))$ by solving $h(x_i \mid \mathrm{pa}(x_i)) - z_i = 0$
    end if
    if $X_i$ is discrete then
        Determine $x_i$ such that $x_i = \min\{x : z_i \le h(x \mid \mathrm{pa}(x_i))\}$
    end if
end for
```

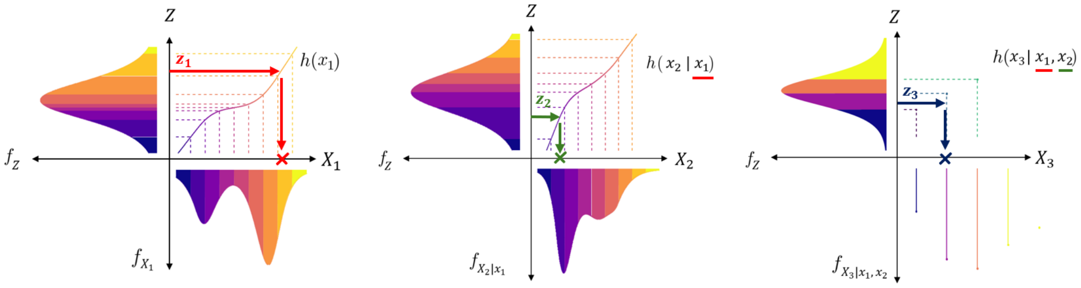


Figure 2.4: One sampling iteration for the three variables from the estimated transformation functions $h(x_i \mid \mathrm{pa}(x_i))$. The latent values z_i are sampled from the standard logistic distribution. The values x_i are determined by applying the inverse of the transformation function for continuous variables or by finding the corresponding category for the ordinal variable.

Interventional sampling To sample from the interventional distribution, we can apply the do-operator as described by Pearl (1995) (Pearl named it set instead of do). The do-operator fixes a variable at a certain value and sample from the distribution of the other variables while keeping the fixed variable constant. For example, if one wants to intervene on X_2 and set it to a specific value α, $\mathrm{do}(x_2 = \alpha)$ and then sample from the interventional-distribution

$$x_3 = \min \{x : z_3 \leq h(x \mid x_1, x_2 = \alpha)\}$$

with the same process as for the observational sampling, with the only difference that the intervened variable X_2 stays constant.

Counterfactual queries In a counterfactual query one wants to know what the value of variable X_i would have been if another variable X_j had a different value than what was acutally observed. Pearl (2009b) describes the three-step process to answer counterfacutal queries as follows: Given a causal model M and observed evidence e (which are the actually observed values of the variables X_i of one sample) one wants to compute the probability of $Y = y$ under the hypothetical condition $X = x$.

Here input image of counterfactuals from final presentation.

Step 1 aims to explain the past (Z) by knowledge of the evidence e; Step 2 amends the past to the hypothetical condition $X = x$ Step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$

Pearl named these three steps, (1) abduction, (2) action and (3) prediction. The procedure is described in the pseudocode **??** and illustrated in Figure.

Algorithm 2 Answer a Single Counterfactual Query

```
Given: A structural model $X_k = f(Z_k, \mathrm{pa}(X_k))$, with inverse noise map $Z_k = h(X_k \mid \mathrm{pa}(X_k))$
Input: Observed sample $x$, intervention $X_i := \alpha$, target variable $X_j$
Step 1: Abduction Infer latent variable $Z_j = h(x_j \mid \mathrm{pa}(x_j))$ using the observed values
Step 2: Action Replace the value of $X_i$ with $\alpha$ in the set of parent variables
Step 3: Prediction Compute the counterfactual value $x_j^{cf} = h_j^{-1}(Z_j \mid \mathrm{pa}(x_j)^{cf})$
```

While the probability of Y under the hypothetical condition $X = x$ can be determined in any case, the actual counterfactual value of Y is only defined for a continuous outcome but not for discrete outcomes.

(What pearl writes: Likewise, in contrast with the potential-outcome framework, counterfactuals in the structural account are not treated as undefined primitives but rather as quantities to be derived from the more fundamental concepts of causal mechanisms and their structure.)

2.2 Individualized Treatment Effect (ITE)

2.2.1 Terminology: ITE vs. CATE

In causal inference, the term *Individual Treatment Effect (ITE)* typically refers to the unobservable difference in potential outcomes for a specific individual:

$$\mathrm{ITE}(x) = Y_i(1) - Y_i(0), \tag{2.7}$$

where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes for individual i under treatment and control, respectively. However, this individual-level effect is inherently unobservable, as we only observe one of the two potential outcomes for each individual.

Instead, most methods aim to estimate the expected treatment effect conditioned on a given set of covariates X. This quantity is known as the *Conditional Average Treatment Effect (CATE)*:

$$\mathrm{CATE}(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}[Y \mid X = x, T = 1] - \mathbb{E}[Y \mid X = x, T = 0]. \tag{2.8}$$

In applied fields such as healthcare, marketing, and machine learning, it is common to refer to estimates of the CATE as *individualized treatment effects (ITE)*. Despite the slight abuse of terminology, this reflects the practical goal of estimating treatment effects tailored to individual characteristics.

Terminology Note: In this thesis, we use the term *ITE* to refer to estimated individualized treatment effects, i.e., estimates of the CATE. This choice is made for consistency with model outputs, figures, and existing terminology in the literature we build upon.

Curth *et al.* (2024) provide a comprehensive overview of the individualized treatment effect (ITE) and its estimation in the context of causal machine learning. They state its importance in comparison to average treatment effects, the assumpitons that need to be fulfilled, what kind of limiations typically are encountered and how models should be validated.

Randomized controlled trials (RCTs) are considered the gold standard for estimating causal effects due to their ability to eliminate confounding through randomization. However, RCTs typically report the *average treatment effect (ATE)*, which summarizes the effect of a treatment across an entire study population. This obscures individual-level variation in treatment response: some individuals may benefit substantially, others not at all, or even be harmed. In personalized medicine and risk-based decision-making, such population-level summaries are insufficient. Instead, the objective is to guide treatment decisions tailored to individual patient characteristics, for which the *individualized treatment effect (ITE)* is a more appropriate target. Where the homogeneous treatment effect refers to the part of the effect that is equal for all patients, the heterogeneous treatment effect describes the non-random variation in treatment effects across individuals or groups.

The Potential Outcomes Framework Causal inference is commonly formalized within the *Rubin Causal Model*, also known as the potential outcomes framework. For each individual i, let $Y_i(1)$ denote the potential outcome under treatment, and $Y_i(0)$ the outcome under control. The individual treatment effect is defined as

$$\tau_i = Y_i(1) - Y_i(0). \tag{2.9}$$

However, only one of the two potential outcomes can be observed for each individual, which constitutes the *fundamental problem of causal inference*. Related estimands include the *conditional average treatment effect (CATE)*,

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x], \tag{2.10}$$

which reflects the expected effect conditional on covariates $X = x$, and the more general concept of *heterogeneous treatment effects (HTE)*.

In contrast to the mean-based estimands above, the *quantile treatment effect (QTE)* evaluates differences in the distribution of potential outcomes. For instance, the median treatment effect is defined as

$$\tau^{(0.5)}(x) = Q_{Y(1)|X=x}(0.5) - Q_{Y(0)|X=x}(0.5), \tag{2.11}$$

where $Q_{Y(t)|X=x}(q)$ denotes the q-th quantile of the potential outcome under treatment t. QTEs are particularly relevant when treatment effects are not symmetrically distributed or when tail behavior is of interest. We will later perform a simulation study using the median for the QTE estimation.

2.2.2 Assumptions for Identifiability

above paper also discusses the implications of strong identifiability assumpions and that cate still can be biased if there are unobserved interaction variables. (however, not sure how good this paper is, many typos)

To identify treatment effects from observational data, several key assumptions are required. *Consistency* ensures that the observed outcome equals the potential outcome under the received treatment. The *Stable Unit Treatment Value Assumption (SUTVA)* assumes no interference between units and that treatments are well-defined. The most critical assumption is *ignorability* (or unconfoundedness), which posits that, conditional on covariates X, the treatment assignment is independent of the potential outcomes:

$$(Y(1), Y(0)) \perp T \mid X. \tag{2.12}$$

In addition, the *positivity* assumption requires that the probability of receiving each treatment is strictly between 0 and 1 for all covariate strata:

$$0 < P(T = 1 \mid X = x) < 1. \tag{2.13}$$

These assumptions are untestable but are necessary for identifying causal effects from non-randomized data.

2.2.3 Propensity Score Adjustment in Observational Settings

In the absence of randomization, the *propensity score*, defined as the probability of receiving treatment conditional on covariates X,

$$e(X) = P(T = 1 \mid X), \tag{2.14}$$

can be used to balance treatment groups and reduce confounding rosenbaum1983central. When the ignorability assumption holds, adjusting for the propensity score allows for unbiased estimation of average treatment effects. However, while propensity score methods (e.g., matching, stratification, weighting) are effective for estimating population-level quantities like the ATE, they are often insufficient for ITE estimation. Since ITE requires modeling both potential outcomes at the individual level, direct modeling approaches such as outcome regression, meta-learners, or Bayesian models are typically more appropriate rubin2007design, nie2021quasi. Moreover, reliance on the propensity score alone may fail to capture fine-grained individual heterogeneity necessary for personalized treatment decisions ali2019addressing.

If ITEs are correct, they should give on average the ATE: E(ITE) = ATE, (cite a paper, maybe Alicia Curth or Hoogland?). However, we can not say that the ITEs must be correct, if they give the ATE on average. (because they could for example also be much more spread than the true ITEs but still give the correct ATE on average!)

Check Hooglands paper (guide) again, he should also say something about benefits of RCT that might be relevant for my subjects.

Rubins potential outcomes framework.

Quantile Treatment effect

- also talk about propensity score Rubin(2007) to basically estimate an RCT...and overcome the problem of confounding. but this might work for ATE but not really for ITE, direct modelling of the outcome is necessary

Models for ITE Estimation

Assessing predictive performance with AUC, calibration slope, and brier score. In Leo presentation, he says that recalibration can either be done with deviance statistics or by leave one out (loo) cross validation the slope of this regression would then be the estimated shrinkage factor. T-learner vs s-learner, metalearner, virtual twins

The ITE for a binary endpoint is estimated as the difference of two probabilities (the risk under treatment minus the risk under control). It is essential that the model used to estimate these probabilities is well calibrated and generalizes to new (unseen) data. (Guo *et al.*, 2017) point out that even though modern neural networks became much more accurate in terms of prediction performance, they are no longer well-calibrated. As it may not be a big problem for the sole purpose of making good predictions, it is very problematic for applications where an accurate quantification of the uncertainty is needed. When using models that are estimated with conventional methods such as ordinary least squares or standard maximum likelihood, they tend to overfit on the training data and make too extreme predictions on the test data. This problem increases with reduced sample size, low event rate or large number of predictor variables. To prevent such overfitting in regression models, penalization (shrinkage) methods are proposed as they shrink the estimated coefficients towards zero to reduce the variance in predictions on new data (Riley *et al.*, 2021).

Logistic regression, penalized logistic regression (shrinkage, lasso Shrinkage methods should provide better predictive perfomance on average (cite articles). Calster *et al.* (2020) analyzed different regression shrinkage methods with a binary outcome in a simulation study. They concluded, although the calibration slope improved on average, shrinkage often worked poorly on individual datasets. With small sample size and low number of events per variable the performance of most of these methods were highly variable and should be used with caution in such settings. Riley *et al.* (2021) obtained to similar results in their simulation study. Problems occur, because tuning parameters are often estimated with large uncertainty on the training data and fail to generalize. In both studies the autors pointed out that these penalization methods are more unreliable when needed most, that is when the risk of overfitting may be large.

When using a Random Forest based method as an S-learner (where Treatment is given as input variable) one has to make sure that the treatment variable is realli used in the forests and not left out by mtry (not all variables are used for splits because decorrelating and reducing overfit.)

(Shrinkage shrinks the coefficients so that the calibration slope is improved on a test set. The shrinkage factor can for example be found with n-fold cross validation, as e.g. done by lasso with L1 penalization)

Explain tuning of random forest, with depth number of trees and mtry (ranger?)

TRAM-DAGs with complex shifts or compmlex intercepts can capture heterogeneity. See appendix XXX for an example of ITE estimation with a complex shift.

Use of instrumental variables (IV) to also estimate CATE in presence of unobserved confounders (Nichols, 2007), (Hartford *et al.*, 2017). Frauen and Feuerriegel (2023) propose a model based on IV that is said to also be applicable on observational data.

Talk about that interaction variables are also referred to effect modifiers, they are no confounders (not necessarily) and also no mediators. An example could be the psychological condition of a patient which might also affect how the treatment works, this is not a confounder but an effect modifier, and i would assume that this variable is rarely recorede or measured.

In this thesis , I will apply Lasso regression on the IST stroke trial and simulation studies, where the sample size is relatively large.

"This could include the analysis of individual patient data from multiple randomized trials, or even the use of nonrandomized studies for the estimation of outcome risk under a control condition." this motivates the need for observational modeling.

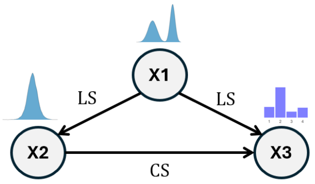
Problems with ITE: (in an RCT setting) - to estimate the ITE we must assume un-confoundedness. Does this also apply to itneractions (effect modifiers)? Check how this is handled in the literature.

– somewhere, explain how Confidence intervals (based on bootstrap or wald CI) were computed, for ITE-ATE plot or other things.

2.3 Experiment 1: TRAM-DAG (simulation study)

In this experiment, we demonstrate the application of TRAM-DAGs on a synthetic dataset, using the same illustrative DAG shown previously in Figure 2.3. The aim is to show how TRAM-DAGs can learn structural equations from observational data, by fitting a model to the joint distribution that resulted from the underlying causal structure. We visualize the model fitting in terms of the training loss, and subsequently show and interpret the learned components of the transformation functions, such as intercepts, linear and complex shifts. Finally, we draw samples from the estimated distributions to get observational and interventional distributions. We also conduct counterfactual queries on the learned model.

Data generating process: We simulate a dataset consisting of three variables, X_1, X_2, and X_3, following the structure of the DAG and its associated meta-adjacency matrix as shown in Figure 2.5. The matrix describes the functional dependencies between the variables, where LS indicates a linear shift and CS a complex shift. Rows denote the source of effect, and columns the target.



$$\longrightarrow \mathbf{MA} = \begin{bmatrix} 0 & \text{LS} & \text{LS} \\ 0 & 0 & \text{CS} \\ 0 & 0 & 0 \end{bmatrix}$$

Figure 2.5: Causal graph (left) and meta-adjacency matrix (right) for experiment 1. The transformation function of X_2 depends on X_1 by a linear shift (LS). The transformation function for X_3 depends on X_1 by a linear shift (LS) and on X_2 by a complex shift (CS).

The variable X_1 is continuous and bimodally distributed, and acts as a source node in the DAG, i.e., it is not influenced by any other variable:

$$X_1 = \begin{cases} \mathcal{N}(0.25,\, 0.1^2) & \text{with probability } 0.5, \\ \mathcal{N}(0.73,\, 0.05^2) & \text{with probability } 0.5 \end{cases}$$

The second variable, X_2, is continuous and linearly dependent on X_1 on the log odds scale, with a true coefficient of $\beta_{12} = 2$. Its transformation function is given by

$$h(X_2 \mid X_1) = h_I(X_2) + \beta_{12} X_1,$$

where the baseseline transformation of X_2 is $h_I(X_2) = 5X_2$.

The third variable, X_3, is ordinal and depends on both X_1 (LS) and X_2 (CS). We define the complex functional shift by X_2 as $f(X_2) = 0.5 \cdot \exp(X_2)$, and the linear shift by X_1 as $\beta_{13} = 0.2$. The transformation for category k of the ordinal variable X_3 with 4 levels (K) is thus defined by

$$h(X_{3,k} \mid X_1, X_2) = \vartheta_k + \beta_{13} X_1 + f(X_2),$$

with cut-points $\vartheta_k \in \{-2,\, 0.42,\, 1.02\}$ defining the thresholds of the ordinal variable. The samples for X_2 and X_3 are generated as described in Section 2.1.4, by first sampling a latent value from the standard logistic distribution, and then determining the corresponding observation according to the transformation function.

This simulation allows us to assess whether the TRAM-DAG model can correctly recover the underlying structural dependencies and parameters (linear and complex).

Model: Given the adjacency matrix and the simulated observations, we construct a modular neural network based on the TRAM-DAG framework. The complex shift from X_2 to X_3 is modeled by a neural network with 4 hidden layers with 2 nodes each, as illustrated in Figure 2.6. 20'000 samples are generated according to the defined DGP for fitting the model. The model is trained for 400 epochs with the Adam optimizer (Kingma and Ba, 2015) and a learning rate of 0.005.

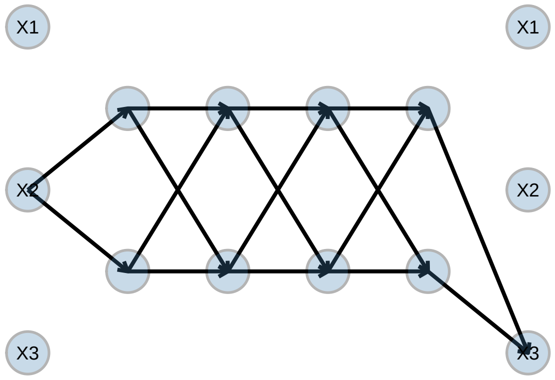


Figure 2.6: Neural network architecture for the complex shift on X_3 from X_2. The complex shift is modelled by a neural network with 4 hidden layers of shape (2, 2, 2, 2).

Evaluation: We compare the estimated intercepts, learned coefficients and the complex shift to the true values used in the DGP. We compare the sampled observational and interventional distributions to the true distributions. For the counterfactual queries, we show the estimated counterfactual values for X_2 when intervening on X_1 at a specific value, and compare them to the true counterfactual values.

2.4 Experiment 2: ITE on International Stroke Trial (IST)

Chen *et al.* (2025) evaluated multiple causal ML methods on the International Stroke Trial (IST), to estimate the individualized treatment effects. They demonstrated that none of the applied ML methods generalized well, as performance on the test data differed significantly from the training data on the chosen evaluation metrics. In this experiment, we replicate the analysis on the same data by applying three causal ML methods for ITE estimation, to investigate whether we obtain similar results as the authors.

Data: The International Stroke Trial was a large, randomized controlled trial conducted in the 1990s to assess the efficacy and safety of early antithrombotic treatment in patients with acute ischemic stroke (International Stroke Trial Collaborative Group, 1997). Using a 2x2 factorial design, 19'435 patients across 36 countries were randomized within 48 hours of symptom onset to receive aspirin, subcutaneous heparin, both, or neither. Patients allocated to aspirin (300 mg daily for 14 days) had a 6-month death or dependency rate of 62.2%, compared to 63.5% in the control group not receiving aspirin, corresponding to a statistically significant absolute risk reduction after adjustment for baseline prognosis (1.4%, $p = 0.03$). The authors stated that there was no interaction between aspirin and heparin in the main outcomes. In this thesis, we focus exclusively on the aspirin vs. no aspirin comparison and the outcome of death or dependency at 6 months after stroke.

The dataset used in this experiment was made publicly available by Sandercock *et al.* (2011) and contains individual-level data, including baseline covariates assessed at randomization, treatment allocation, and 6-month outcomes, with a follow-up rate of 99%.

We use the same data pre-processing steps as Chen *et al.* (2025) to ensure comparability of results. 5.9% of individuals had incomplete data and were removed from the dataset. We use 2/3 of the data for fitting the models and 1/3 as a hold out test set. The final dataset included 21 baseline variables recorded at randomization: aspirin allocation (treatment), age, delay between stroke and randomization (in hours), systolic blood pressure, sex, CT performed before randomization, visible infarct on CT, atrial fibrillation, aspirin use within 3 days prior to randomization, and presence or absence of neurological deficits (including face, arm/hand, leg/foot deficits, dysphasia, hemianopia, visuospatial disorder, brainstem or cerebellar signs, and other neurological deficits), as well as consciousness level, stroke subtype, and geographical region. The outcome variable was death or dependence at 6 months.

Models for ITE estimation: The aim is to estimate the ITE based on the baseline characteristics. As benchmark, we apply a T-learner logistic regression (as in Chen *et al.* (2025)). As a representation of a more complex model we apply a T-learner tuned random forest (comets) and finally a S-learner TRAM-DAG. For the random forest and TRAM-DAG based methods, we additionally scale numerical and dummy encode categorical covariates prior to model training. The transformation for the outcome is modelled by a complex intercept $h(Y \mid T, X) = CI(T, X)$, with 4 hidden layers of shape (20, 10, 10, 2). This architecture allows for interaction between the treatment and covariates. Furthermore, batch normalization, ReLU activation, and dropout (0.1) are applied to prevent overfitting and stabilize learning. A validation set comprising 20% of the training data is used to select the model with the lowest out-of-sample negative log-likelihood (NLL), while the test set remains untouched for final evaluation. Since the IST stroke trial is a randomized controlled trial, the full potential of TRAM-DAGs (that lies in the observational setting) is not needed, as only the outcome has to be modelled as a function of the baseline patient characteristics. Nevertheless, this is not a reason not to apply it.

Model evaluation: For validation, since the ground truth is not known, we first rely on calibration plots to assess the general prediction power for the probabilities. Second, we predict the potential outcomes with the trained models to estimate the ITE on the training and test set in terms of the risk difference $\text{ITE}_i = \text{P}(Y_i = 1|T = 1, X_i) - \text{P}(Y_i = 1|T = 0, X_i)$. For visual validation, we will show the densities of the estimated ITEs on both datasets and the ITE-ATE plots to assess whether the estimated ITEs align with the actually observed treatment effects

2.5 Experiment 3: ITE model robustness under RCT conditions (simulation study)

In this section, we will perform a simulation study to estimate the ITE with different models in an RCT setting under different scenarios. The aim is to identify conditions under which ITE estimation fails and whether such failure is model-agnostic – i.e., driven by external factors such as unobserved covariates or treatment effect magnitude, rather than by the model class itself. It may provide insight into why ITE estimation can fail in a real-world application as, for example, demonstrated by Chen *et al.* (2025) on the IST stroke trial and replicated in our own work in Section 2.4. The simulation is based on a data generating process (DGP) that should resemble an RCT. We assume a binary outcome and a set of covariates that influence the treatment effect. There may also be treatment-covariate interactions that are responsible for heterogeneity of treatment effect.

Data generating process: The data is generated similarly as proposed by Hoogland *et al.* (2021). The binary treatment (T) was sampled from a Bernoulli distribution with probability 0.5. The 5 covariates ($\mathbf{X}$) represent patient specific characteristics at baseline and were drawn from a multivariate standard normal distribution with a compound symmetric covariance matrix ($\rho = 0.1$). The binary outcome (Y) is sampled from a Bernoulli distribution with probability $\text{P}(Y_i = 1 \mid \mathbf{X_i} = \mathbf{x_i}, T_i = t_i) = \text{logit}^{-1}\left(\beta_0 + \beta_T t_i + \boldsymbol{\beta}_X^\top \mathbf{x_i} + t \cdot \boldsymbol{\beta}_{TX}^\top \mathbf{x_{TX,i}}\right)$, where i denotes the patient indicator and $\mathbf{x}_{TX,i}$ denotes the subset of covariates that interact with the treatment. The data is generated for three scenarios, where the coefficients are set to different values or not all variables are observed. In scenario 1, the covariates are set as follows: $\beta_0 = 0.45$ (intercept), $\beta_T = -0.85$ (direct treatment effect), $\boldsymbol{\beta}_X = (-0.5, 0.8, 0.2, 0.6, -0.4)$ (direct covariate effects), and $\boldsymbol{\beta}_{TX} = (0.9, 0.1)$ (interaction effects between treatment and covariates X_1 and X_2 on the outcome). In scenario 2, the same coefficients are used but the covariate X_1, which is responsible for a large portion of the heterogeneity, is not observed in the final dataset. This is expected to cause difficulties in estimating the ITE. In scenario 3, the coefficients for the direct treatment and interaction effects are set to $\beta_T = -0.05$ and $\boldsymbol{\beta}_{TX} = (-0.01, 0.03)$ to represent a weak treatment effect and low heterogeneity, all other coefficients stay unchanged and all variables are observed. The DAGs of the three scenarios are presented in Figure 2.7.

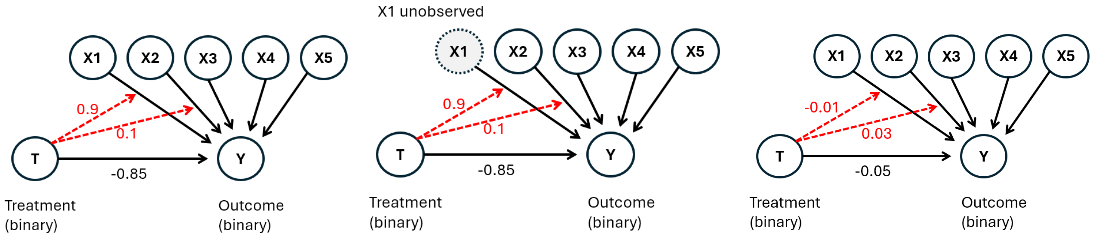


Figure 2.7: Data generating process (DGP) for the three scenarios in the ITE simulation study (RCT). Interaction effects between treatment (T) and covariates (X_1 and X_2) on the outcome (Y) are shown in red. Left: DGP for scenario 1, where all covariates are observed and there is a strong treatment effect and heterogeneity; Middle: DGP for scenario 2, with same DGP as in scenario 1, but where the covariate X_1 is not observed; Right: DGP for scenario 3, where the treatment effect and heterogeneity is weak and all covariates are observed.

Models for ITE estimation: The datasets generated from the DGP under the three scenarios are used to estimate the ITE with different models. We applied the following models: T-learner logistic regression (`stats`-package), T-learner logistic lasso regression (`glmnet`-package (Friedman *et al.*, 2010), regularization parameter lambda estimated with 10-fold cross-validation), S-learner logistic lasso regression (same as T-learner), T-learner random forest (`randomForest`-package (Breiman, 2001), 100 trees), T-learner tuned random forest (`comets`-package, tunes the number of variables to possibly split at in each node (mtry) and the maximal tree depth (max.depth) parameters out-of-bag, 500 trees). While all models were applied, we will present only the results of the T-learner logistic regression as benchmark (same model as used in the data generating process), and the tuned random forest as representation of a complex non-parametric model. In the Appendix 6.7 we further present the results for a default random forest evaluated on scenario 1, to show the importance of tuning the model, which means to prevent overfitting and ensure accurate calibration. All models were trained on a training set and evaluated on a test set with 10'000 samples each, generated from the same DGP. TRAM-DAGs would also be well suited for ITE estimation in this setting, but we chose not to apply it in this experiment, since the main objective is to assess behavioral differences between complex and simple models under different scenarios. TRAM-DAGs are applied in the other experiments in this thesis.

Model evaluation: The model results are evaluated visually for the training and test dataset. For the predictive performance, we present the true vs. predicted probabilities P(Y=1| X, T) which should give an impression on how well the model is calibrated. Plots of the true vs. predicted ITE show how close the model predictions are to the true effects. The true probabilities and ITEs are known by design in this simulation, allowing for direct assessment of calibration, which would not be available in a real world-application. Finally, we present the ITE-ATE plot, which shows whether the estimated ITEs correspond to the actual observed outcomes, and the discrepancy between training and test set. The observed ATE in terms of risk difference $\text{ATE} = \text{P}(Y = 1|T = 1) - \text{P}(Y = 1|T = 0)$ is calculated and plotted for each ITE subgroup. Because the ITE is defined here as the difference between the predicted outcome probabilities under treatment and control, it is also expressed on the risk-difference scale. Following, the

ITE-ATE plot can be understood as a calibration plot, where an ideal model should represent the identity line.

Whether the estimated ITEs correspond to the actual observed outcomes, and the discrepancy between training and test set, is assessed with the ITE-outcome plot. These simulation scenarios allow us to assess how ITE estimation performance behaves under challenging conditions such as omitted variables and weak effect size. The subsequent results reveal which models are robust to such violations and provide insight into real-world estimation failures.

2.6 Experiment 4: ITE estimation with TRAM-DAGs (simulation study)

We claim that if the assumptions for ITE estimation (identifiability, (unobserved) unconfoundedness etc.) are fulfilled and the DAG is fully known, with the TRAM-DAG framework, the ITE can be estimated under observational data just like with RCT data. We aim to proof this with the DAG as displayed in Figure 2.8. The binary treatment (X4) is the intervention variable and we want to estimate the ITE for the continuous outcome Y.

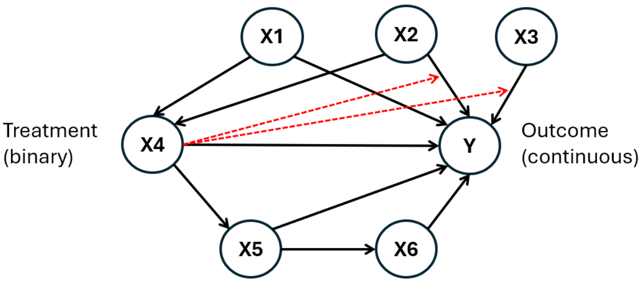


Figure 2.8: DAG used for the experiment to estimate the ITE. DGP: the source nodes X1, X2 and X3 come from a multivariate standard normal distribution with 0.1 correlation. In the observational setting the binary treatment X4 depends on the parents X1 and X2, in the RCT Setting, this dependency is omitted due to randomization. X5 depends on the treatment X4. X6 depends on X5. The outcome Y depends on all variables with additional interaction effects between the treatment and the variables X2 and X3. All variables except the treatment X4 are continuous.

An example scenario that would have the structure of the proposed dag could be the following: A marketing campaign is conducted to increase customer spending. The treatment is the marketing email (X4) that is sent to the customers. If the treatment is not randomized, it depends on the prior total spend (X1) and the customer engagement score (X2). The outcome is the total spend in the next 30 days after receiving the email (X7). The prior total spend and customer engagement score are confounders that influence both the treatment and the outcome. Customer satisfaction score (X3) from a recent survey is another predictor. The time spent on the website after receiving the email (X5) is a mediator that influences the number of product pages viewed (X6), which in turn influences the total spend in the next 30 days.

Data generating mechanism: The standard logistic was chosen as the noise distribution to align with other examples in this thesis. Also any other noise distribution cold be chosen, as we are not interested in interpretability of the coefficients in this experiment. All variables except the binary treatment X4 are continuous. The source nodes X1, X2 and X3 are generated from a multivariate standard normal distribution, where each pair of variables has a correlation of 0.1. These variables represent baseline patient characteristics. In the observational setting, X1 and X2 act as confounders by influencing the treatment allocation X4 and the outcome Y. In the RCT setting, these connections are cut due to randomization. X5 depends on the treatment

X4. X6 depends on X5. The log odds for the continuous outcome are linearly depend on all covariates including additional interaction terms between the treatment and X2 and X3. Hence, the log odds of the outcome can be written in terms of a transformation model with linear shift $h(y \mid X) = h_I(y) + \text{LS}$. Equation 2.15 outlines the outcome on the log odds scale

$$h(y \mid \mathbf{X}) = h_I(y) + \boldsymbol{\beta}_X^\top \mathbf{X} + \boldsymbol{\beta}_{TX}^\top \mathbf{X}_{\text{int}} X_4 \tag{2.15}$$

where $h_I(y)$ is the intercept function, $\mathbf{X}$ is the covariate vector including all variables and $\mathbf{X}_{\text{int}} = \{X2, X3\}$ is the vector with the interaction variables that only has an effect if the treatment is present $(X4 = 1)$. The intercept $h_I(y)$ has to be a smooth monotonically increasing and function which we defined as $h_I(y) = tan(y/2)/0.2$ in the interval between -2 and 2 and linearly extrapolated the function at the boundaries. The coefficients $\boldsymbol{\beta}_X$ are set to $\boldsymbol{\beta}_X = (-0.5, 0.5, 0.2, 1.5, -0.6, 0.4)$, where 1.5 is the direct effect of the treatment X4 on the outcome. $\boldsymbol{\beta}_{TX}$ are set to $\boldsymbol{\beta}_{TX} = (-0.9, 0.7)$ for the interaction terms.

Experiment: The experiment is conducted with 3 different scenarios of data generating mechanism for the outcome Y accordingly: (1) direct and interaction effect, (2) only direct effect, (3) only interaction effect. Depending on the scenario, the corresponding coefficients in $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_{TX}$ are set to zero. The data is generated with a sample size of 20'000 samples for the training set. In both settings, observational and RCT, the TRAM-DAG is first fitted on the data. To allow for full flexibility, all nodes that depend on some parents are modelled by complex intercepts with 3 hidden layers of shape (10, 10, 10). Batch normalization, dropout (0.1) and ReLU activation are used. The model is fitted on the training data consisting of 20'000 samples. To prevent overfitting, an additional validation set with 10'000 samples is used and the model is selected, where the validation loss was is (early stopping).

Once the model is fitted, we obtained the estimated (inverse) transformation functions $X_i = h^{-1}(Z_i \mid pa(X_i))$ that represent the equations $X_i = f(Z_i, pa(X_i))$ in the structural causal model. The process for the ITE estimation is outlined in 3. In a first step to estimate the ITE, we determine the latent values z_{ij} in all observed samples j for the explanatory nodes i - X1, X2, X3, X5 and X6. The latent values are the values of the transformation functions at the observed value of the variable given the observed values of its parents $z_{ij} = h_i(x_{ij} \mid pa(x_{ij}))$. In a second step, these latent values z_{ij} are used to sequentially sample from the two interventional distributions when setting the treatment X4 to either 0 or 1. For each individual, these interventions impact the mediator nodes X5 and X6 as well as the outcome Y. The source nodes X1, X2 and X3 are the same under both treatments. The treatment X4 is the variable which we fix by the do-intervention. X5 and X6 will change according to the treatment. Finally, for each set of samples j (meaning for each individual) we get two distributions for the outcome, one under treatment and one under control. In contrast to the potential outcomes framework, where the potential outcomes are defined as the expected value of the outcome under treatment, we define the potential outcomes as the median of the outcome distribution under treatment - the quantile treatment effect (QTE). For simplicity, we will further refer to the individual treatment effect as ITE even though technically, the QTE is meant. Determining the potential outcomes in terms of the expected values would also be possible, but would require us to repeatedly sample from each resulting potential outcome distribution for each individual and average the results. This was computationally too time consuming and therefore we decided to estimate the QTE instead. In the ITE estimation in the previous examples with binary outcome, this was not necessary, since the potential outcomes were defined as the probabilities of the outcome under treatment and control, hence a single number that represents the expected value.

Notes after Meeting 24.06.25: Depending on the problem, CATE in terms of expected values of potential outcomes might be more appropriate than QTE, but also QTE could be better. Depends. If we wanted the potential outcomes based on the expected values, we have two options. either sample latent values and evaluate inverse tranformation functions. from those two sample distributions calculate the means to get the expected potential outcomes. Lucas suggested that we could also use numerical integreation instead, then we would not have to sample.

Maybe visualize the potential outcome transformation funcitons (both funcitons in one plot) and then show that the median Latent value 0 creates the two potential median outcomes on the x axis.

NOTE: in both, the RCT and in the Observational setting, also other models could be applied instead of TRAM-DAG. As long as all confounders are included in the model, we controll for the confounders and can get unbiased results. For example a T-learner Colr($Y \sim X_1 + X_2 + X_3$) (because Colr is basically what we did in the DGP) fitted on both treatment groups separately could be used to estimate the ITE in our proposed experiment. This might only be possible so easily as long as we do not assume additional interactions between the treatment and the mediators X_5 and X_6. If we would assume such interactions, we would have to include these in the model as well, which would make it more complex and possibliy requires to fit and apply multiple models. If there are no interactions with the mediators, they can be omitted, since we are interested in the total treatment effects and not in separating the effect (mediation analysis). But again, we can only omit if these variables do not contain additional information about treatment effect heterogeneity. The reasoning is because to estimate the total effect one should not control for mediators. (check if really true!!!) However, the TRAM-DAG framework is well suited to also deal with mediators and calculate counterfactuals, therefore we think it is a good example to show its capabilities.

Algorithm 3 ITE Estimation (QTE) Using TRAM-DAG in Observational Data

```
Input: Fitted TRAM-DAG, observational dataset with $n$ samples
for each sample $j = 1$ to $n$ do
    Step 1: Encode explanatory nodes
    for each explanatory node $X_i \in \{X_1, X_2, X_3, X_5, X_6\}$ do
        Compute latent value: $z_{ij} = h_i(x_{ij} \mid \mathrm{pa}(x_{ij}))$
    end for
    Step 2: Generate potential outcomes under treatment and control
    for $x_4 \in \{0, 1\}$ do                    ▷ Simulate both treatment states
        Fix $X_4 = x_4$ (intervention)
        Sample $X_5$ and $X_6$ sequentially using $z_{ij}$ and inverse transformations
        Sample potential outcome $y_j^{(x_4)}$ using $z_{7,i} = 0$ (median of the potential outcome distribution)
    end for
    Step 3: Compute QTE for individual $j$
    $\mathrm{ITE}_j = \mathrm{median}(y_j^{(1)}) - \mathrm{median}(y_j^{(0)})$
end for
Output: ITE estimates $\{\mathrm{ITE}_j\}_{j=1}^n$
```

Validation of results: In the data generating mechanism, along with the actually sampled values, the potential values under both treatments are also recorded and used to determine the true QTE (the ITE based on the 50 percent quantiles of the potential outcome distributions of each individual.) The results are displayed by densities of the estimated ITE, the scatterplots of the true vs. estimated ITE, the ITE-ATE plot with the difference in medians as ATE within subgroups to make it comparable to the estimated ITEs. Furthermore the average of all estimated and true (dgp) ITEs are presented in a table (XX) which should be an estimator (?) for the ATE. We further calculate the ATE as the overall difference in medians in the RCT setting and compare it to the estimated values based on the ITEs. If these estimates are comparable, it would support our claim that with TRAM-DAGs it does not matter if the data is from an RCT or observational setting, as long as the assumptions are fulfilled and the DAG is fully known and observed.

2.7 Software

check report, how i cited the packages. calibration plot, all packages for tram dags, ite, plotting

Maybe it is the methods section. Here however, we give a couple hints. Note that you can wisely use `preamble`-chunks. Minimal, is likely:

Chapter 3

Results

3.1 Experiment 1: TRAM-DAG (simulation study)

In this section, we present the results of a simulation study to evaluate the performance of the TRAM-DAG model in a simple scenario as illustrated by the DAG in Figure 2.5. The model was fitted on synthetic data. Figure 3.1 shows the loss and the estimated parameters for the linear shifts over epochs during training. The loss is minimized during training and the estimated parameters β_{12} and β_{13} converge to the true values used in the DGP. The linear shift parameters are interpretable part of the model (log odds ratios). From the fitted model, we generated samples from the observational distribution, as shown in Figure 3.2. Then we drew samples from the interventional distribution, where $X_2 = 1$ is fixed, as shown in Figure 3.3. Fixing X_2 leads to a distributional change in X_3. The TRAM-DAG model learns the linear shifts (β_{12}, β_{13}) and the complex shift (CS(X_2)), which are shown in Figure 3.4. Figure 3.5 presents the intercepts learned for each of the nodes, with the estimates by the continuous outcomes logistic regression (Colr() function from tram-package (Hothorn *et al.*, 2018)) function as comparison for the continuous variables and the true values used in the DGP for ordinal variable X3 (3 cut points needed for the 4 levels). Finally, Figure 3.6 shows the counterfactuals estimated by the TRAM-DAG model for varying values of X_1. The counterfactuals are the predicted values of X_2 if X_1 would have taken other values instead of the observed value.

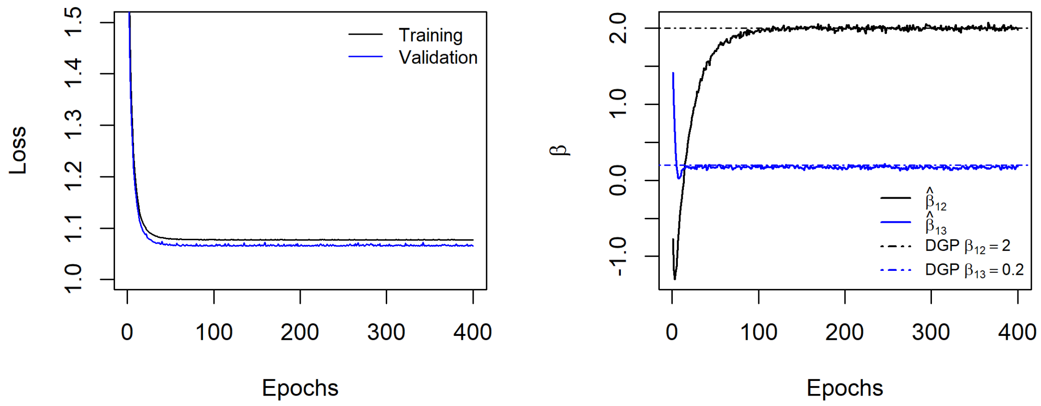


Figure 3.1: TRAM-DAG model fitting over 400 epochs for experiment 1. Left: Loss functions on the training set and a separate validation set; Right: Estimated parameters (betas) for the linear shift components over epochs. They converge to the true values.

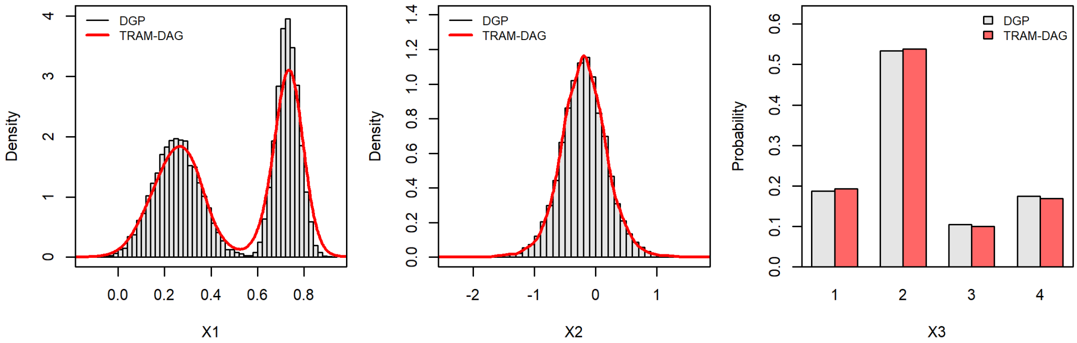


Figure 3.2: Samples by the TRAM-DAG generated from the learned observational against the true observations from the DGP.

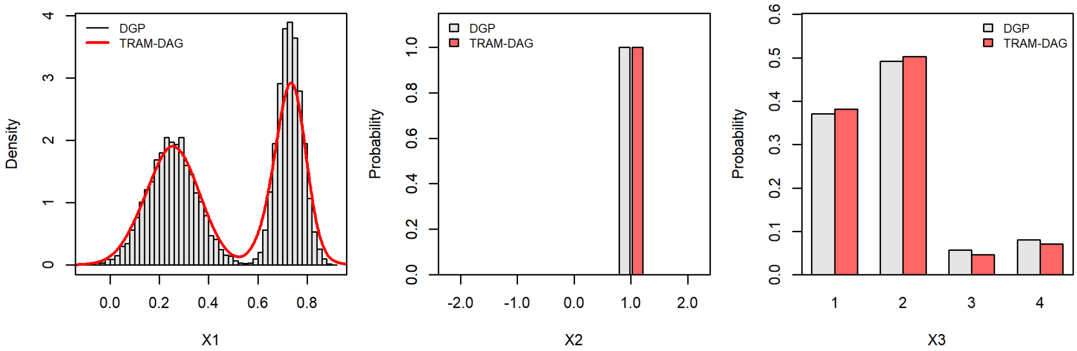


Figure 3.3: Samples by the TRAM-DAG generated against the true observations from the interventional distribution, where $X_2 = 1$ is fixed. According to the DAG, this affects a distributional change in X_3.

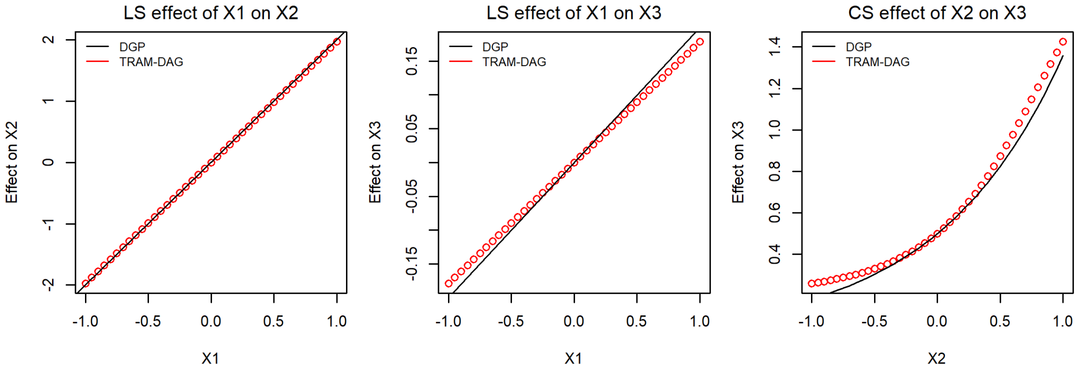


Figure 3.4: Linear shift and complex shift learned by the TRAM-DAG. Left: $\text{LS}(X_1)$ on X_2; Middle: $\text{LS}(X_1)$ on X_3; Right: $\text{CS}(X_2)$ on X_3. For visualization, we subtracted $\delta_0 = \text{CS}(0) - f(0)$ from the estimated complex shift CS(X2) to make it comparable to the DGP shift $f(X_2)$

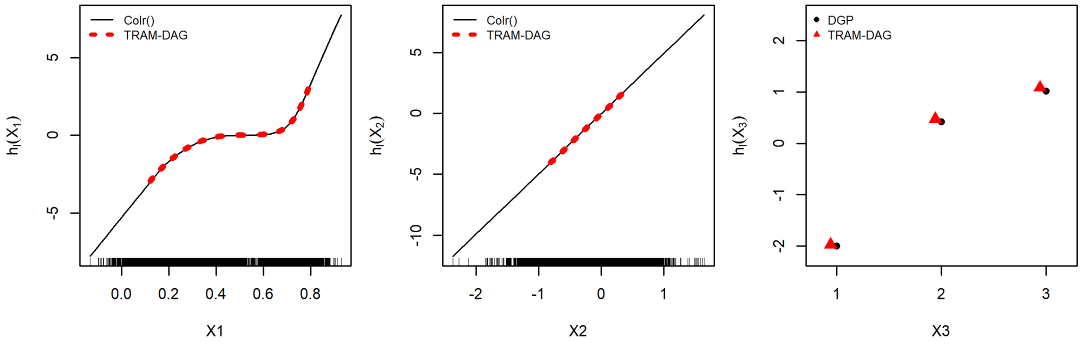


Figure 3.5: Intercepts learned for each of the nodes, with the estimates by the Colr() function for the continuous variables and the true values used in the DGP for ordinal X3. Left: smooth baseline transformation function for continuous X1; Middle: smooth baseline transformation function for continuous X2 ; Right: cut-points as the baseline transformation function for ordinal X3. For the last plot we added $\delta_0 = \text{CS}(0) - f(0)$ to the estimated cut-offs to make them comparable to the true parameters from the DGP.

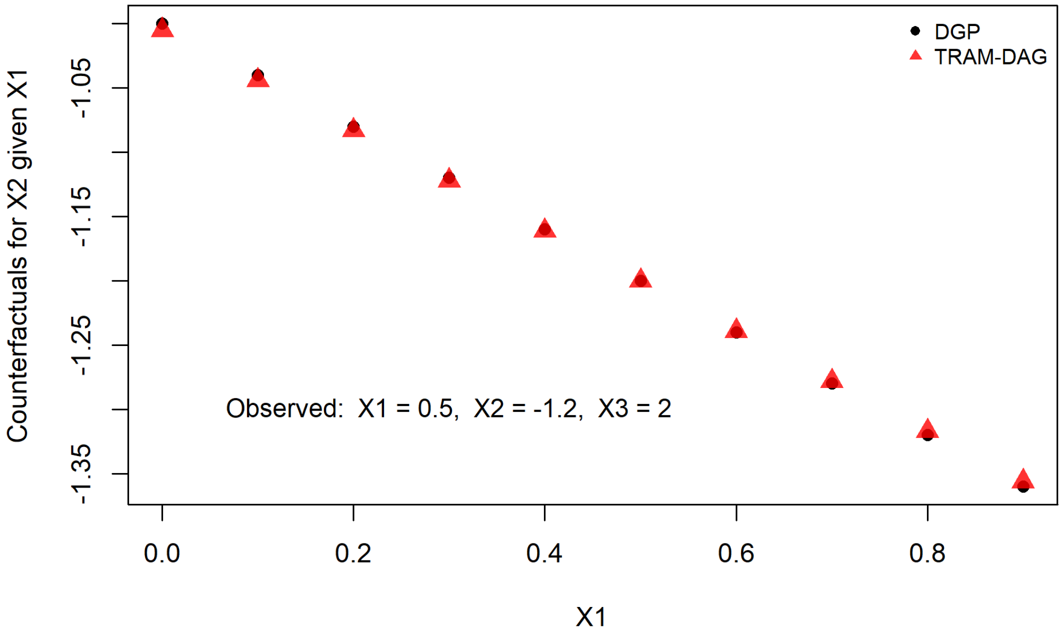


Figure 3.6: Counterfactuals for X_2 estimated with the TRAM-DAG for varying X_1. We assumed observations $X_1 = 0.5$, $X_2 = -1.2$, $X_3 = 2$ and determined the counterfactual values for X_2 if X_1 would have taken other values instead of the observed value.

3.2 Experiment 2: ITE on International Stroke Trial (IST)

In this section, we present the results of the ITE estimation on the International Stroke Trial (IST) dataset. The observed treatment effect $\mathrm{P}(Y = 1|T = 1) - \mathrm{P}(Y = 1|T = 0)$ on the training set was -2.4% absolute risk reduction with a 95% confidence interval of -4.1% to -0.6%. The observed treatment effect on the test set was -0.1% with a 95% confidence interval of -2.6% to 2.3%. The estimated ITEs were computed using three different models: the T-learner logistic regression, the T-learner tuned random forest, and the S-learner TRAM-DAG. The estimated average treatment effect on the test set as $\mathrm{ATE}_{\mathrm{pred}} = \mathrm{mean}(\mathrm{ITE}_{\mathrm{pred}})$ was -2.5% for the T-learner logistic regression, -2.2% for the T-learner tuned random forest, and -3.1% for the S-learner TRAM-DAG. The density of estimated ITEs and ITE-ATE plots in terms of risk difference per estimated ITE subgroup are presented in Figures 3.7 - 3.9. Calibration plots are shown in the Appendix 6.6, Figures 6.1 - 6.3.

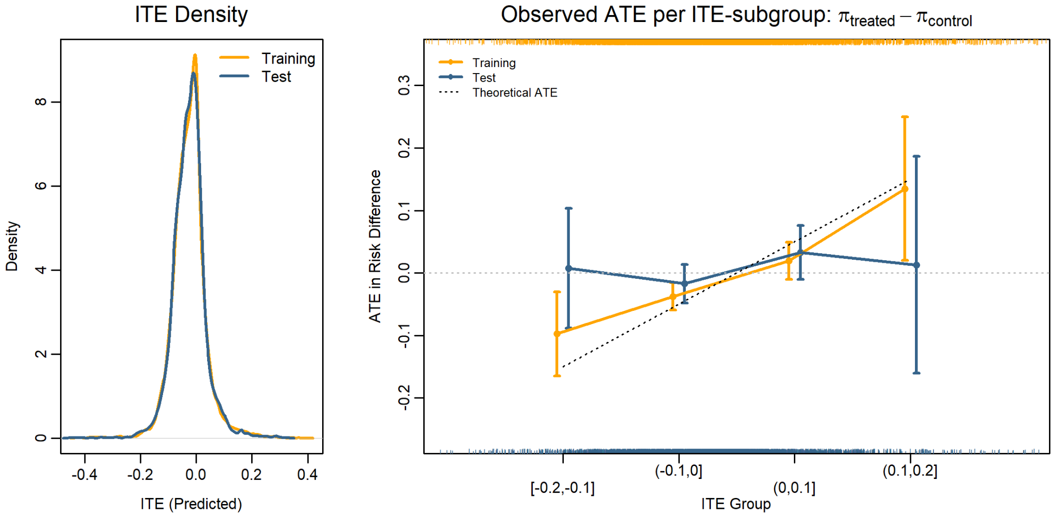


Figure 3.7: Results for the International Stroke Trial (IST) with the T-learner logistic regression. Left: density of the predicted ITE in the training and test set; Right: observed ATE in terms of risk difference per estimated ITE subgroup.

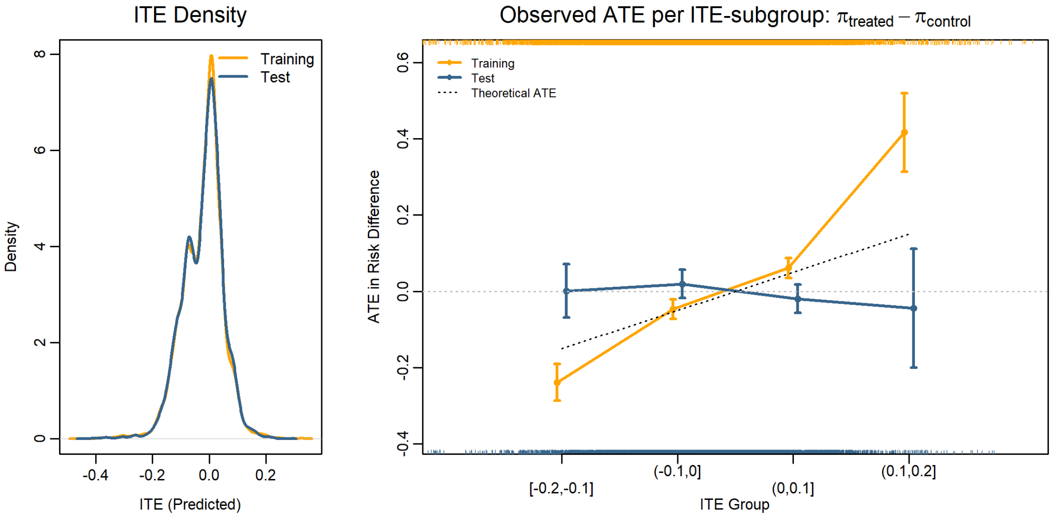


Figure 3.8: Results for the International Stroke Trial (IST) with the T-learner tuned random forest. Left: density of the predicted ITE in the training and test set; Right: observed ATE in terms of risk difference per estimated ITE subgroup.

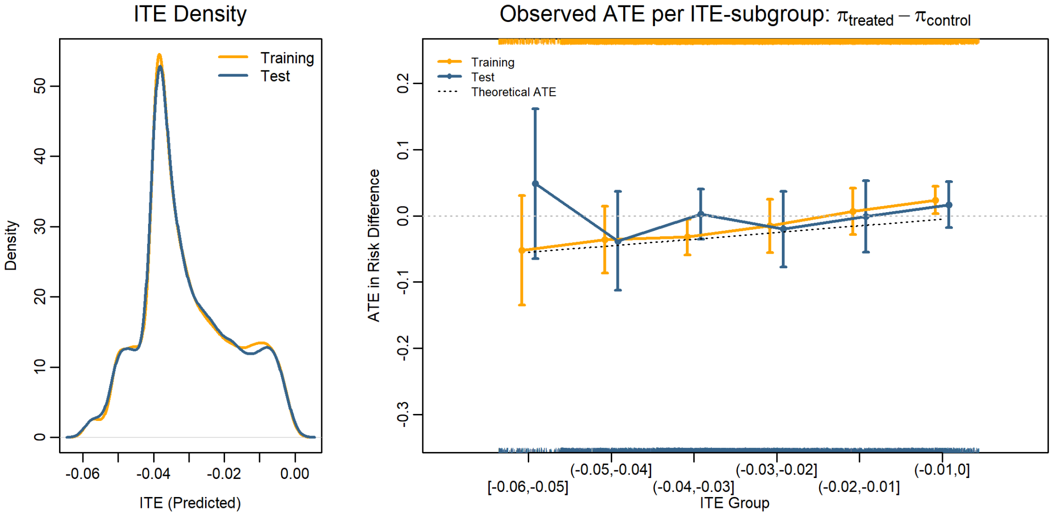


Figure 3.9: Results for the International Stroke Trial (IST) with the S-learner TRAM-DAG. Left: density of the predicted ITE in the training and test set; Right: observed ATE in terms of risk difference per estimated ITE subgroup.

3.3 Experiment 3: ITE model robustness under RCT conditions (simulation study)

In this section, we present the performance of two causal ML models for estimating the ITE under different scenarios. Scenario 1 represents the ideal case where all variables are observed and treatment effects and heterogeneity are large. Scenario 2 uses the same DGP as in scenario 1 but removes the covariate X_1, which has a strong interaction effect with the treatment, from the dataset and treat it as unobserved. Finally, for scenario 3 the coefficients for the direct and interaction treatment effects are weakened, so that heterogeneity is low. All variables are observed again in the last scenario. In each scenario, we applied the T-learner logistic regression and the T-learner tuned random forest. The results of the models on the three scenarios are presented in Figures 3.11 to 3.18.

3.3.1 Scenario (1): Fully observed, large effects

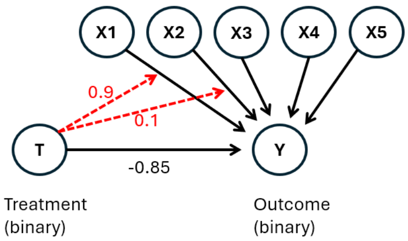


Figure 3.10: DAG for scenario (1), where all variables are observed and there are strong treatment and interaction effects. The numbers indicate the coefficients on the log-odds-scale. Red: interaction effects between treatment (T) and covariates (X_1 and X_2) on the outcome (Y).

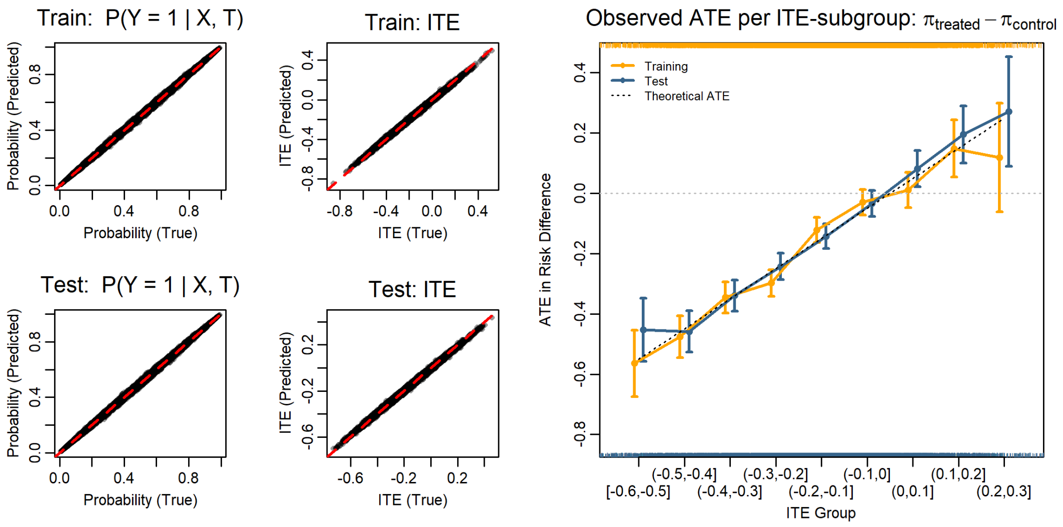


Figure 3.11: Results with the T-learner logistic regression in scenario (1) when the DAG is fully observed and there are strong treatment and interaction effects. Left: true vs. predicted probabilities for $\mathrm{P}(Y = 1 \mid X, T)$; Middle: true vs. predicted ITEs; Right: observed ATE in terms of risk difference per estimated ITE subgroup.

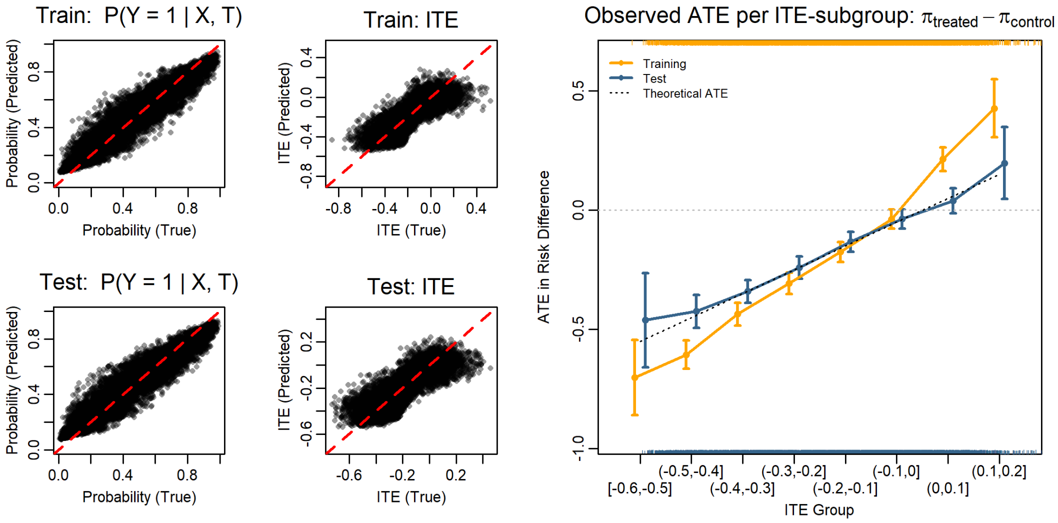


Figure 3.12: Results with the T-learner tuned random forest in scenario (1) when the DAG is fully observed, strong effects. Left: true vs. predicted probabilities for $\mathrm{P}(Y = 1 \mid X, T)$; Middle: true vs. predicted ITEs; Right: observed ATE in terms of risk difference per estimated ITE subgroup.

3.3.2 Scenario (2): unobserved interaction

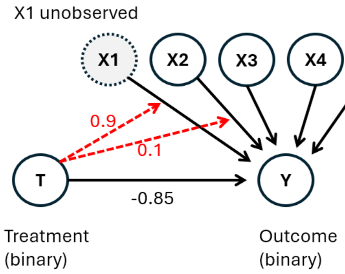


Figure 3.13: DAG for scenario (2), where there are strong treatment and interaction effects, but variable $X1$ is not observed. The numbers indicate the coefficients on the log-odds-scale. Red: interaction effects between treatment (T) and covariates (X_1 and X_2) on the outcome (Y).

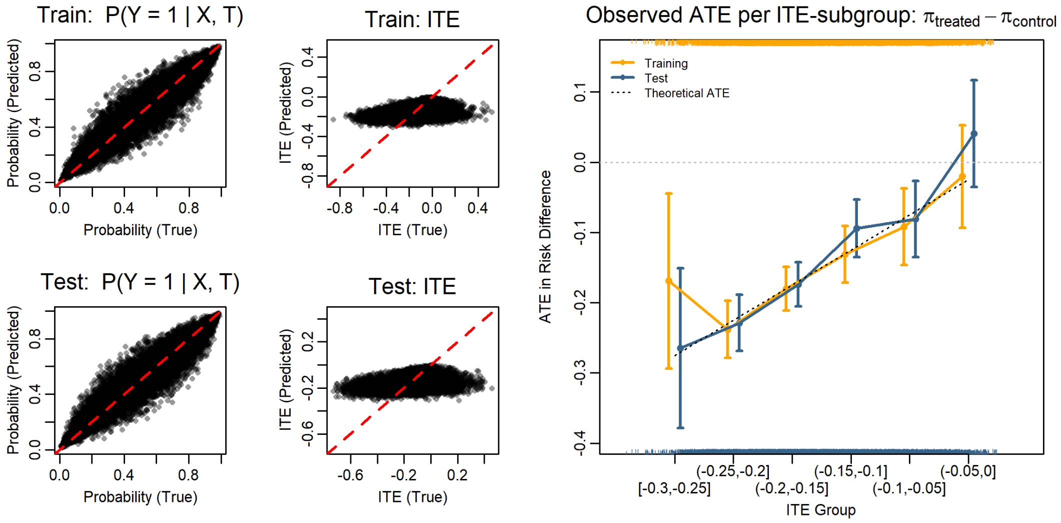


Figure 3.14: Results with the T-learner logistic regression in scenario (2) when there are strong treatment and interaction effects, but variable X_1 is not observed. Left: true vs. predicted probabilities for $\mathrm{P}(Y = 1 \mid X, T)$; Middle: true vs. predicted ITEs; Right: observed ATE in terms of risk difference per estimated ITE subgroup.

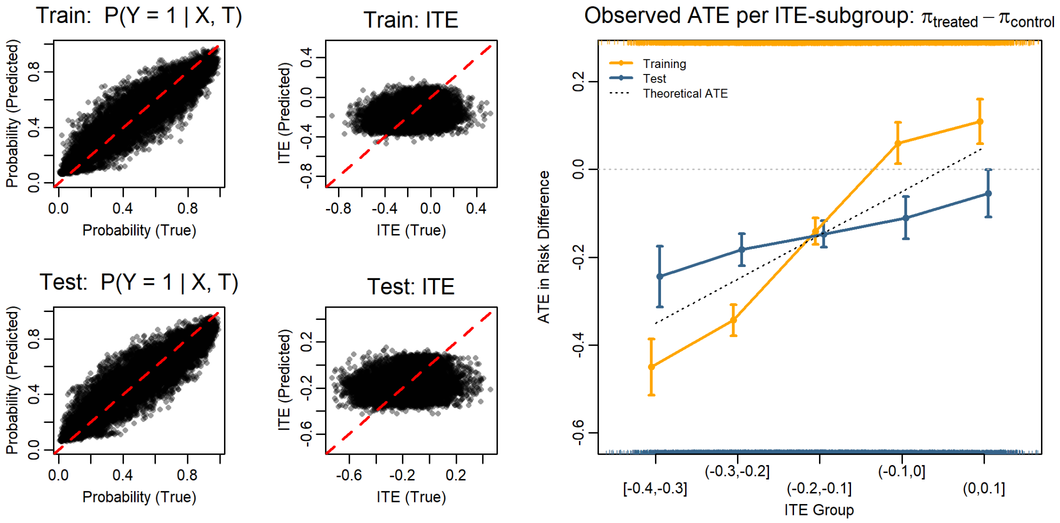


Figure 3.15: Results with the T-learner tuned random forest in scenario (2) when there are strong treatment and interaction effects, but variable X_1 is not observed. Left: true vs. predicted probabilities for $P(Y = 1 \mid X, T)$; Middle: true vs. predicted ITEs; Right: observed ATE in terms of risk difference per estimated ITE subgroup.

3.3.3 Scenario (3): Fully observed, small effects

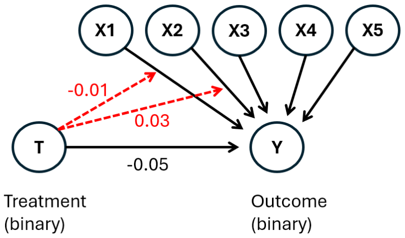


Figure 3.16: DAG for scenario (3), where all variables are observed and there are weak treatment and interaction effects. The numbers indicate the coefficients on the log-odds-scale. Red: interaction effects between treatment (T) and covariates (X_1 and X_2) on the outcome (Y).

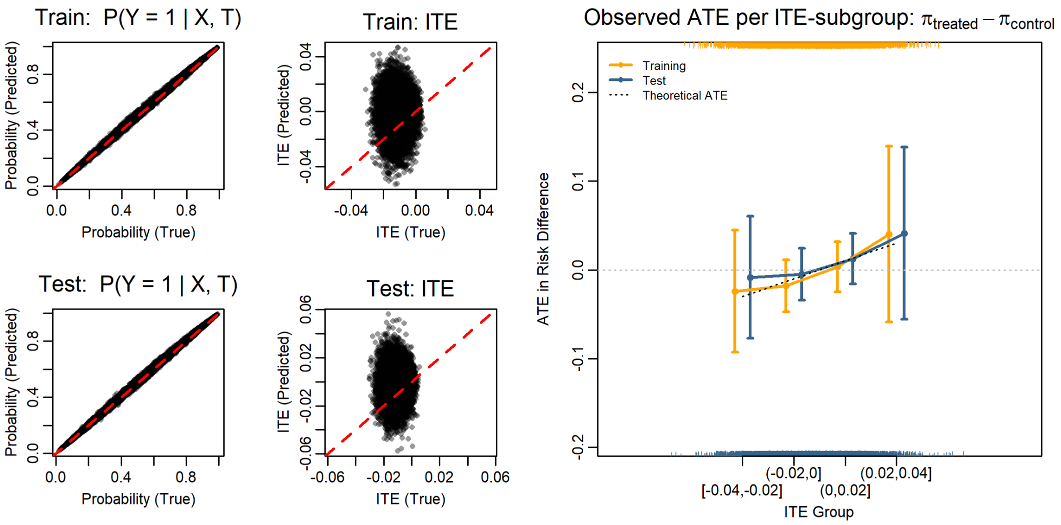


Figure 3.17: Results with the T-learner logistic regression in scenario (3) when the DAG is fully observed and there are weak treatment and interaction effects. Left: true vs. predicted probabilities for $\mathrm{P}(Y = 1 \mid X, T)$; Middle: true vs. predicted ITEs; Right: observed ATE in terms of risk difference per estimated ITE subgroup.

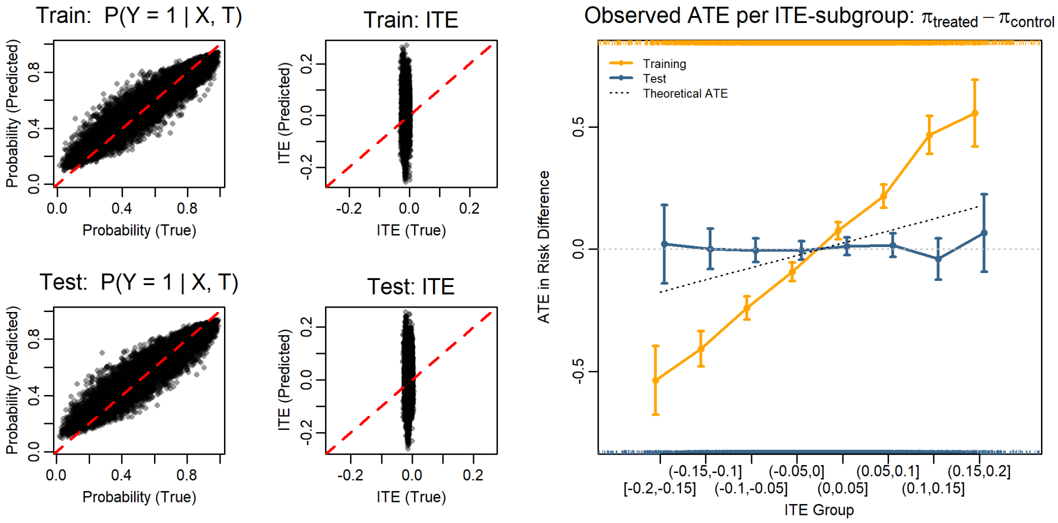


Figure 3.18: Results with the T-learner tuned random forest in scenario (3) when the DAG is fully observed and there are weak treatment and interaction effects. Left: true vs. predicted probabilities for $P(Y = 1 \mid X, T)$; Middle: true vs. predicted ITEs; Right: observed ATE in terms of risk difference per estimated ITE subgroup.

3.4 Experiment 4: ITE estimation with TRAM-DAGs (simulation study)

First, we present the results for scenario (1) with a direct and interaction effect. Then, we present the results for scenario (2) with a direct effect but no interaction effects, and finally, scenario (3) with interaction effects but no direct effect of the treatment. For each scenario, we compare the results in an observational setting with confounded treatment allocation and in a randomized controlled trial (RCT) setting without confounders. We also compare the average treatment effect (ATE), which can directly be calculated in the RCT, with the ATE based on the estimated individualized treatment effects. If the estimated ITEs are unbiased, they should be a good estimate of the ATE. All ITEs presented in this section are technically quantile treatment effects (QTEs) based on the 0.5-quantile of the potential outcomes. For simplicity we will refer to them as ITEs in the following.

3.4.1 Scenario (1): Direct and interaction effects

Scenario (1) included a direct effect of the treatment on the outcome and an additional interaction effect of the treatment with the covariates X2 and X3. A train and test set were generated with 20'000 observations each. In the observational setting, the treatment allocation was confounded by the covariates X1 and X2. In the train set, 38.6% of patients were in the control group and 61.4% were in the treatment group. This ratio was similar in the test set. In the RCT setting treatment allocation was randomized. In the train set 49.8% individuals were in the control group and 50.2% in the treatment group. In the test set 50.2% were in the control group and 49.8% in the treatment group. Figure 3.19 illustrates the true ITE distribution that resulted from the DGP. Due to the interaction effects, there is some heterogeneity in the ITE distribution. Figure 3.20 shows the marginal distributions of all variables according to the DGP and the estimates of the fitted TRAM-DAG. Figure 3.21 shows the distribution of the outcome under the do(Tr=0) and do(Tr=1) interventions. The fitted model was applied to estimate the ITEs in terms of the difference in medians of the potential outcomes. The resulting density of the estimated ITEs compared to the true ITEs according to the DGP is shown in Figure 3.22. Across both settings, the densities of the estimated ITEs are close to the true densities in both the training and test datasets. Figure 3.23 shows the scatterplots of true against estimated ITEs. Finally, Figure 3.24 displays the ITE-ATE plot where the ATE is computed as the difference in medians of the observed outcome under the treatments within the respective ITE-subgroups The trends observed in the training and test sets are consistent.

The average treatment effect (ATE) is presented in Table 3.1. In the RCT setting in the training set, the difference in means of the outcomes in the two treatment groups was -0.563 with a confidence interval of -0.582 to -0.543. The ATE in terms of the difference in medians of the observed outcomes was -0.626. Also in the training set, the ATE in terms of the mean of the true ITEs was -0.62 and the ATE in terms of the mean of the estimated ITEs was -0.619. All measures, including the ones from the test datasets, are shown in Table 3.1.

NOTE: also add CIs in the table with the ATEs?

Table 3.1: Scenario (1), including direct and interaction effects: Comparison of ATE measures across train and test sets for the observational and RCT setting.

Measure	**Observational**		**RCT**	
	Train	**Test**	**Train**	**Test**
ATE as mean($\mathrm{Y}^{(1)}_{\text{observed}}$) $-$ mean($\mathrm{Y}^{(0)}_{\text{observed}}$)	NA	NA	-0.563	-0.563
ATE as median($\mathrm{Y}^{(1)}_{\text{observed}}$) $-$ median($\mathrm{Y}^{(0)}_{\text{observed}}$)	NA	NA	-0.626	-0.638
ATE as mean($\mathrm{ITE}_{\text{true}}$)	-0.62	-0.622	-0.62	-0.622
ATE as mean($\mathrm{ITE}_{\text{estimated}}$)	-0.617	-0.62	-0.619	-0.622

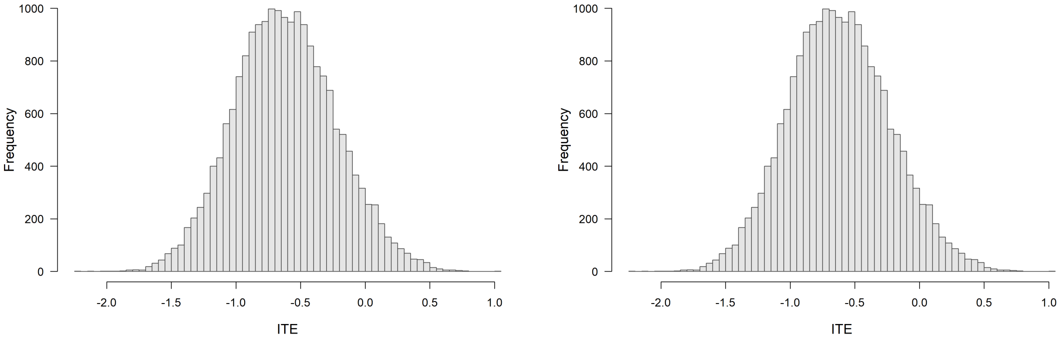


Figure 3.19: True ITE distribution resulting from the DGP for scenario (1) with direct and interaction effects. The true ITEs are identical in the observational and in the RCT setting, since they depend on the potential outcomes under both treatment allocations. Left: Observational; Right: RCT setting.

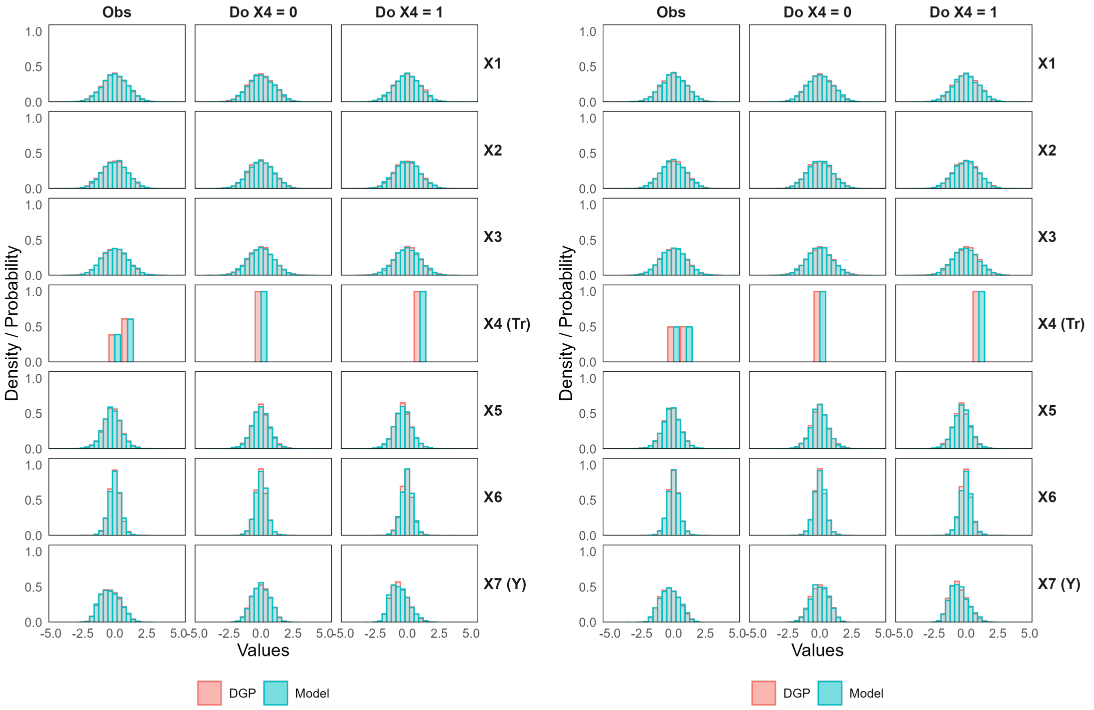


Figure 3.20: Marginal distributions of DGP variables and fitted TRAM-DAG samples for scenario (1) with direct and interaction effects. The distributions shown as observed (Obs), under control intervention (Do $X4 = 0$) and under treatment intervention (Do $X4 = 1$). Left: Observational; Right: RCT setting.

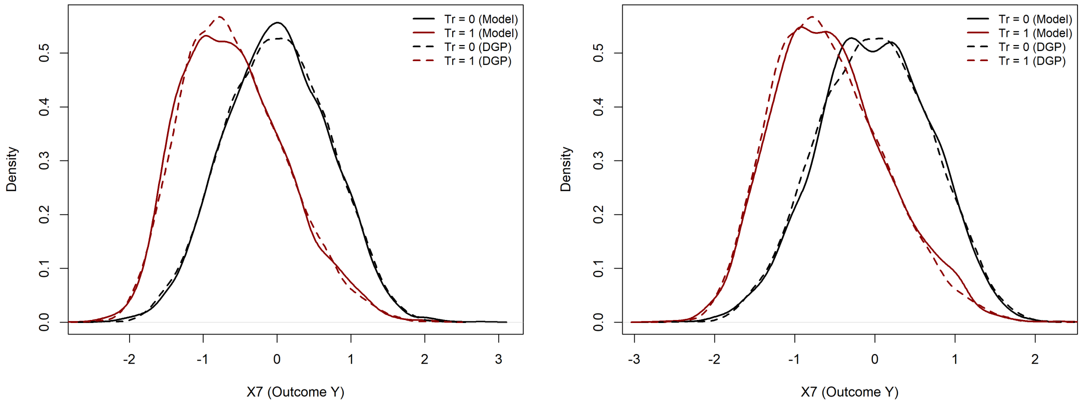


Figure 3.21: Distributions of the outcome variable (X7) under treatment and control interventions for scenario (1), including direct and interaction effects. This plot is a higher resolution view of the X7 panels (Do $X4 = 0$) and (Do $X4 = 1$) from Figure 3.20. Left: Observational; Right: RCT setting.

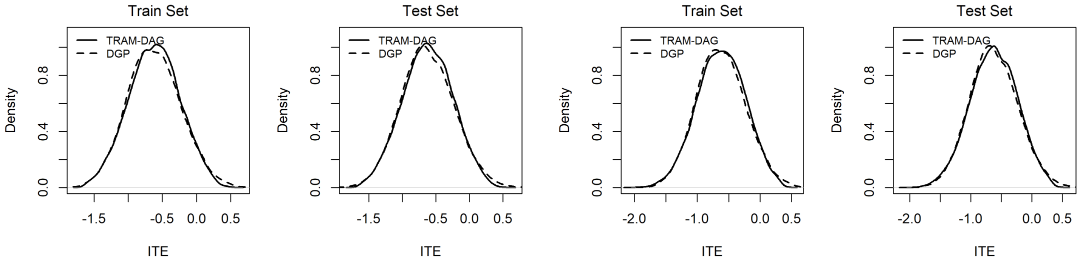


Figure 3.22: Densities of estimated ITEs compared to the true ITEs in the training and test datasets for scenario (1), including direct and interaction effects. Left: Observational; right: RCT setting.

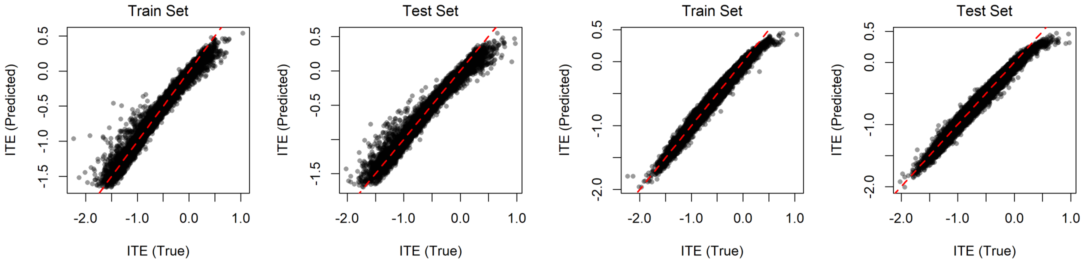


Figure 3.23: Scatterplots of estimated ITEs compared to the true ITEs in the training and test datasets for scenario (1), including direct and interaction effects. Left: Observational; right: RCT setting.

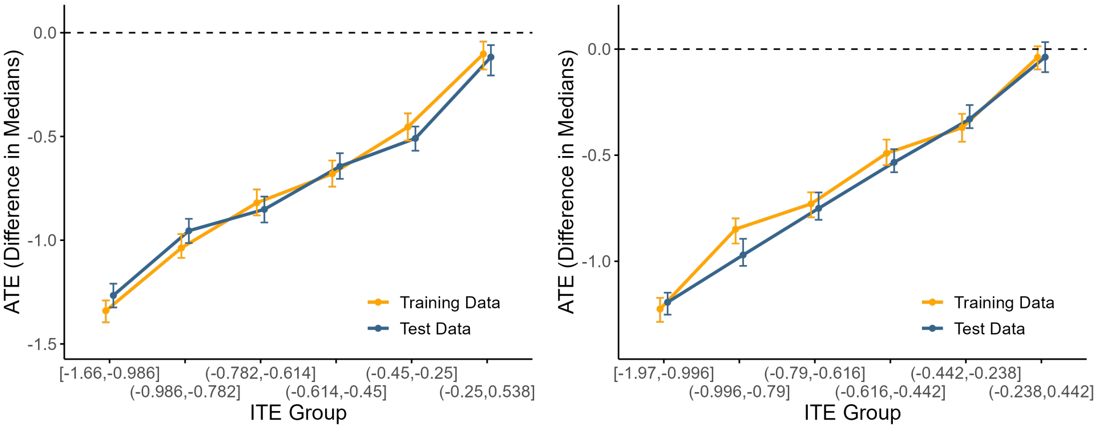


Figure 3.24: ITE-ATE plot for scenario (1), including direct and interaction effects. Individuals are grouped into bins according to the estimated ITE and in each bin the ATE is calculated as the difference in medians of the observed outcomes under the treatments. 95% bootstrap confidence intervals indicate the uncertainty. Left: Observational; right: RCT setting.

3.4.2 Scenario (2): With direct but no interaction effects

Scenario (2) included a direct effect of the treatment on the outcome and coefficients of the interaction effects are set to zero. This results in less heterogeneity of ITE compared to scenario (1) as shown in Figure 3.25. The observational and interventional densities sampled by the fitted TRAM-DAG are aligned with the true densities according to the DGP as illustrated in Figures 3.26 and 3.27. A notable discrepancy in variance exists between the estimated and true ITEs, as illustrated in Figures 3.28 and 3.29. The ITE-ATE plot in Figure 3.30 shows a less informative view compared to scenario (1). Table 3.2 presents the ATE measures for scenario (2). In the test set of the RCT setting, the ATE in terms of the difference in medians of the observed outcomes was -0.639. In contrast, the ATE based on the estimated ITEs in the same dataset was -0.586.

Table 3.2: Scenario (2), including a direct treatment but no interaction effects: Comparison of ATE measures across train and test sets for the observational and RCT setting.

Measure	**Observational**		**RCT**	
	Train	**Test**	**Train**	**Test**
ATE as mean($\mathrm{Y}^{(1)}_{\mathrm{observed}}$) − mean($\mathrm{Y}^{(0)}_{\mathrm{observed}}$)	NA	NA	-0.569	-0.572
ATE as median($\mathrm{Y}^{(1)}_{\mathrm{observed}}$) − median($\mathrm{Y}^{(0)}_{\mathrm{observed}}$)	NA	NA	-0.629	-0.639
ATE as mean($\mathrm{ITE}_{\mathrm{true}}$)	-0.633	-0.633	-0.633	-0.633
ATE as mean($\mathrm{ITE}_{\mathrm{estimated}}$)	-0.645	-0.644	-0.587	-0.586

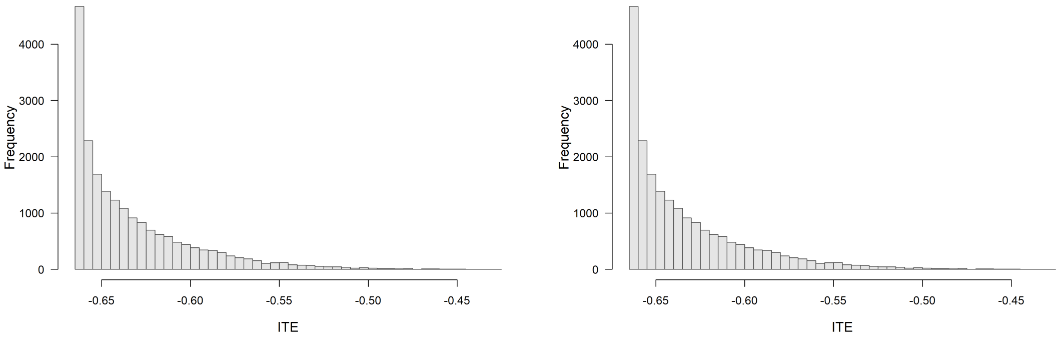


Figure 3.25: True ITE distribution resulting from the DGP for scenario (2), including a direct treatment but no interaction effects. The true ITEs are identical in the observational and in the RCT setting, since they depend on the potential outcomes under both treatment allocations. Left: Observational; Right: RCT setting.

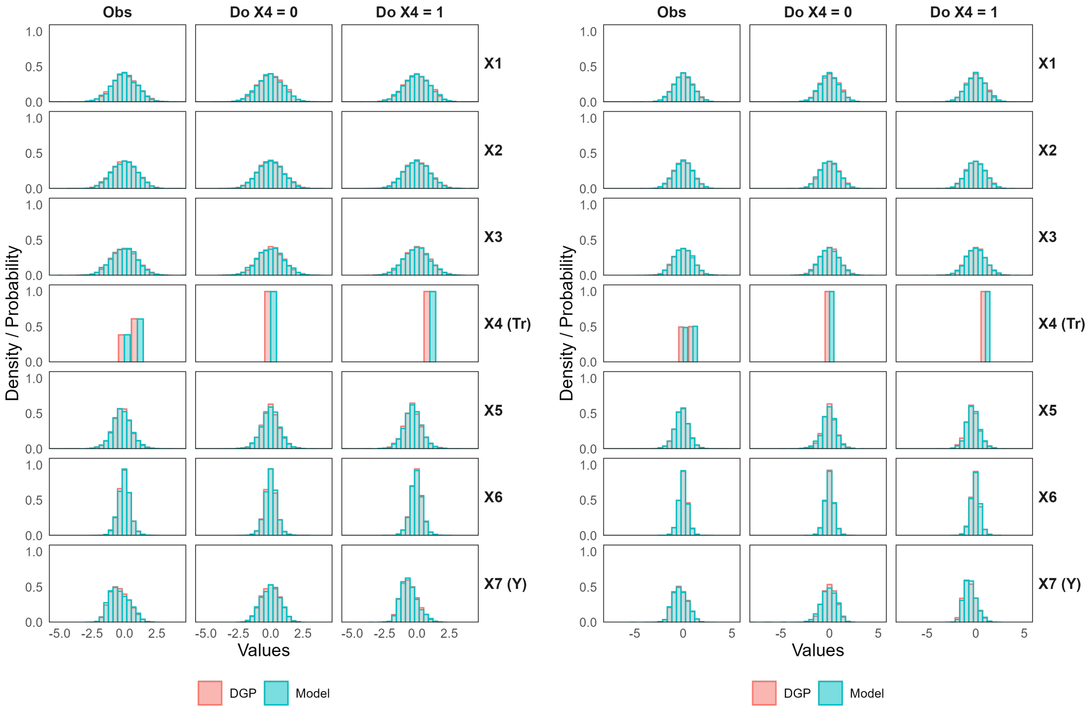


Figure 3.26: Marginal distributions of DGP variables and fitted TRAM-DAG samples for scenario (2), including a direct treatment but no interaction effects. The distributions shown as observed (Obs), under control intervention (Do $X4 = 0$) and under treatment intervention (Do $X4 = 1$). Left: Observational; Right: RCT setting.

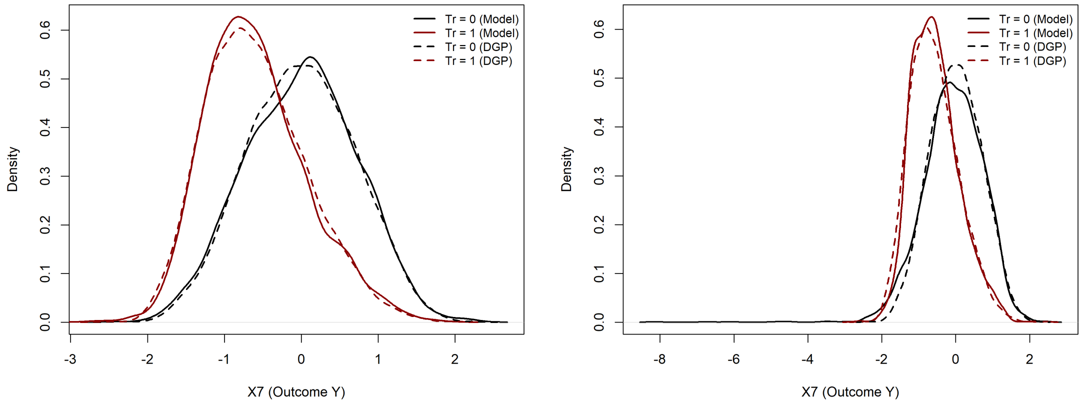


Figure 3.27: Distributions of the outcome variable (X7) under treatment and control interventions for scenario (2), including a direct treatment but no interaction effects. This plot is a higher resolution view of the X7 panels (Do $X4 = 0$) and (Do $X4 = 1$) from Figure 3.26. Left: Observational; Right: RCT setting.

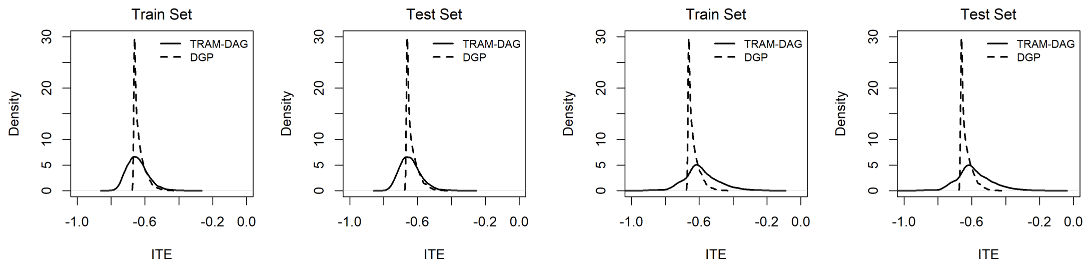


Figure 3.28: Densities of estimated ITEs compared to the true ITEs in the training and test datasets for scenario (2), including a direct treatment but no interaction effects. Left: Observational; right: RCT setting.

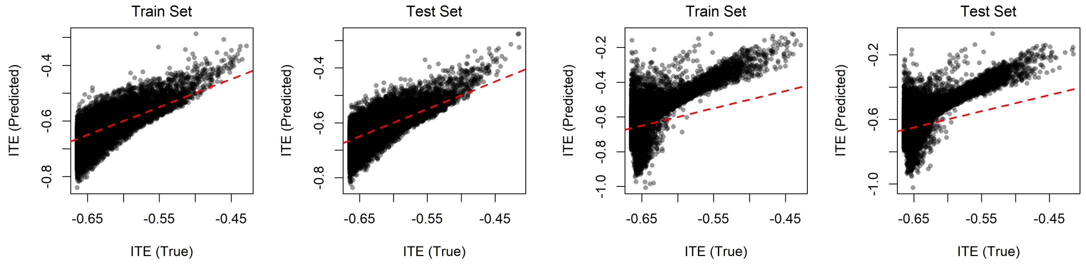


Figure 3.29: Scatterplots of estimated ITEs compared to the true ITEs in the training and test datasets for scenario (2), including a direct treatment but no interaction effects. Left: Observational; right: RCT setting.

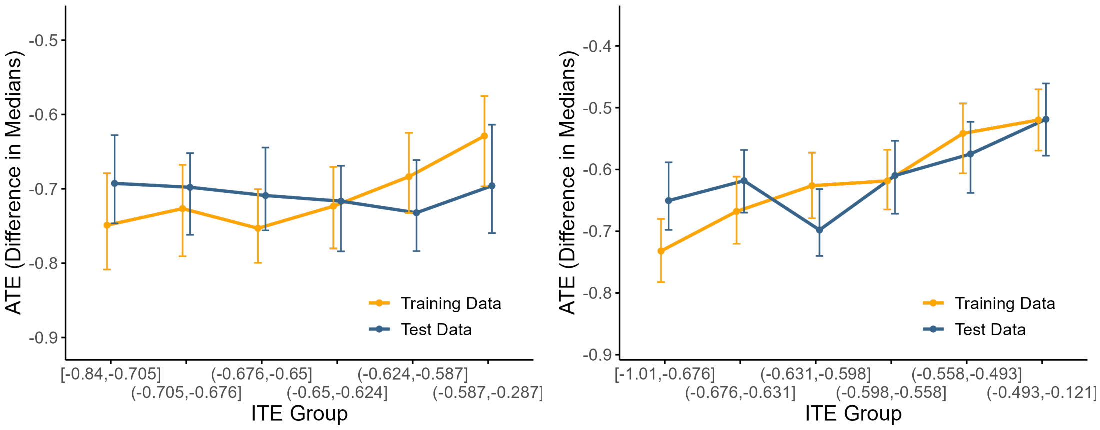


Figure 3.30: ITE-ATE plot for scenario (2), including a direct treatment but no interaction effects. Individuals are grouped into bins according to the estimated ITE and in each bin the ATE is calculated as the difference in medians of the observed outcomes under the treatments. 95% bootstrap confidence intervals indicate the uncertainty. Left: Observational; right: RCT setting.

Â¨

3.4.3 Scenario (3): No direct but with interaction effects

Scenario (3) included no direct effect of the treatment on the outcome but it included interaction effects of the treatment with the covariates X2 and X3. Compared to scenario (1), when excluding the direct effect of the treatment, the distribution of ITEs is more centered as shown in Figure 3.31. The ATE in terms of the mean difference in the test set of the RCT setting is -0.048 with a confidence interval of -0.068 to -0.028.

Table 3.3: Scenario (3), without direct treatment effect but including interaction effects: Comparison of ATE measures across train and test sets for the observational and RCT setting.

Measure	**Observational**		**RCT**	
	Train	**Test**	**Train**	**Test**
ATE as mean($\mathrm{Y}^{(1)}_{\mathrm{observed}}$) $-$ mean($\mathrm{Y}^{(0)}_{\mathrm{observed}}$)	NA	NA	-0.048	-0.048
ATE as median($\mathrm{Y}^{(1)}_{\mathrm{observed}}$) $-$ median($\mathrm{Y}^{(0)}_{\mathrm{observed}}$)	NA	NA	-0.048	-0.059
ATE as mean($\mathrm{ITE}_{\mathrm{true}}$)	-0.065	-0.068	-0.065	-0.068
ATE as mean($\mathrm{ITE}_{\mathrm{estimated}}$)	-0.059	-0.061	-0.051	-0.053

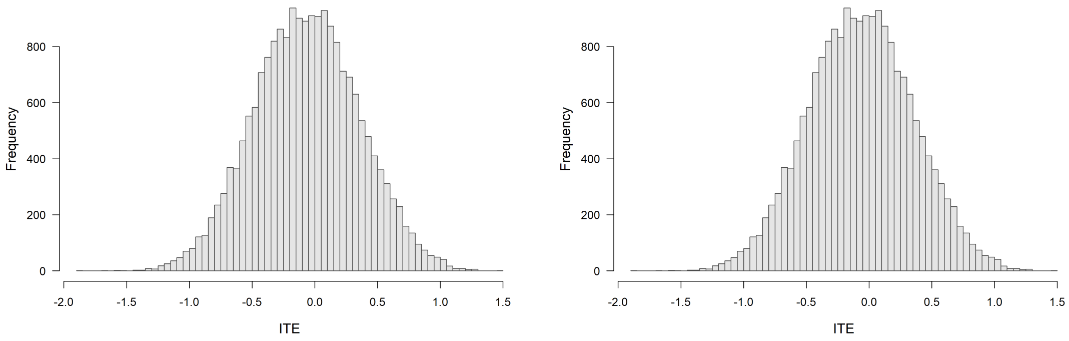


Figure 3.31: True ITE distribution resulting from the DGP for scenario (3), without direct treatment effect but including interaction effects. The true ITEs are identical in the observational and in the RCT setting, since they depend on the potential outcomes under both treatment allocations. Left: Observational; Right: RCT setting.

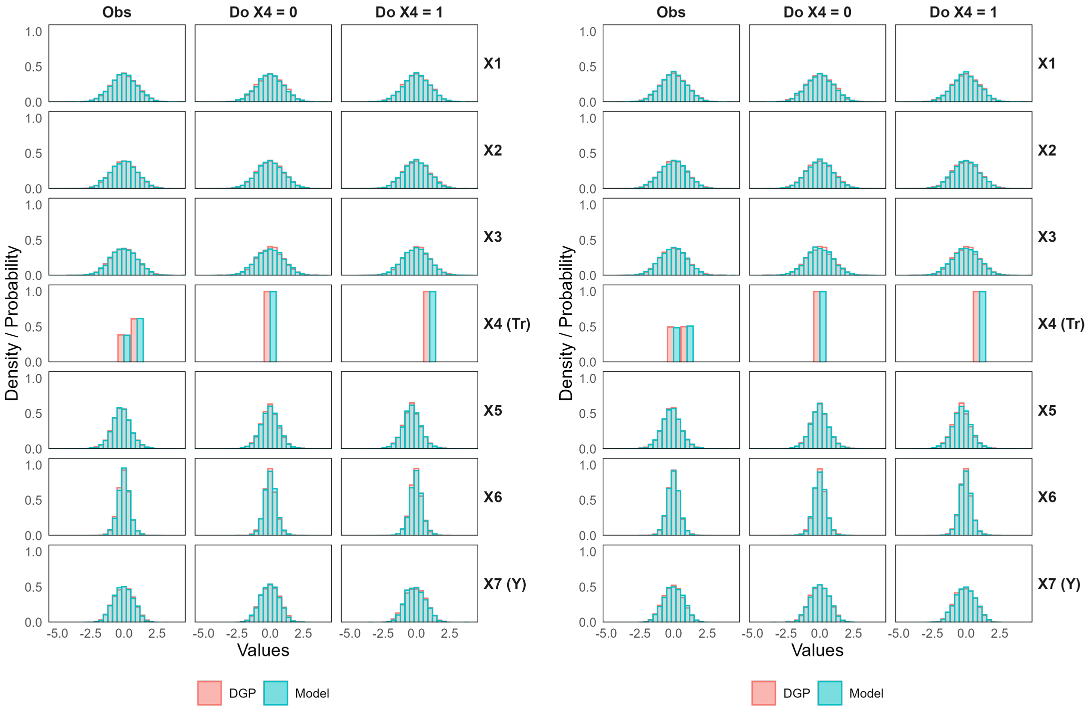


Figure 3.32: Marginal distributions of DGP variables and fitted TRAM-DAG samples for scenario (3), without direct treatment effect but including interaction effects. The distributions shown as observed (Obs), under control intervention (Do $X4 = 0$) and under treatment intervention (Do $X4 = 1$). Left: Observational; Right: RCT setting.

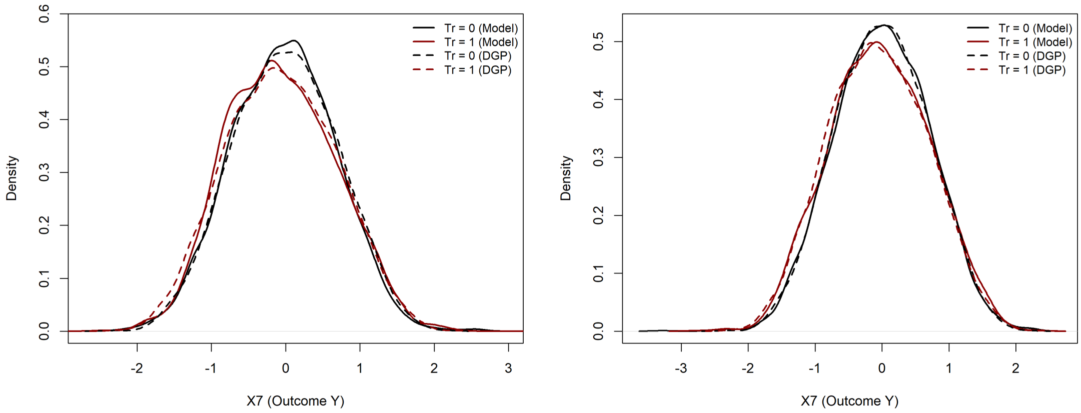


Figure 3.33: Distributions of the outcome variable (X7) under treatment and control interventions for scenario (3), without direct treatment effect but including interaction effects. This plot is a higher resolution view of the X7 panels (Do $X4 = 0$) and (Do $X4 = 1$) from Figure 3.32. Left: Observational; Right: RCT setting.

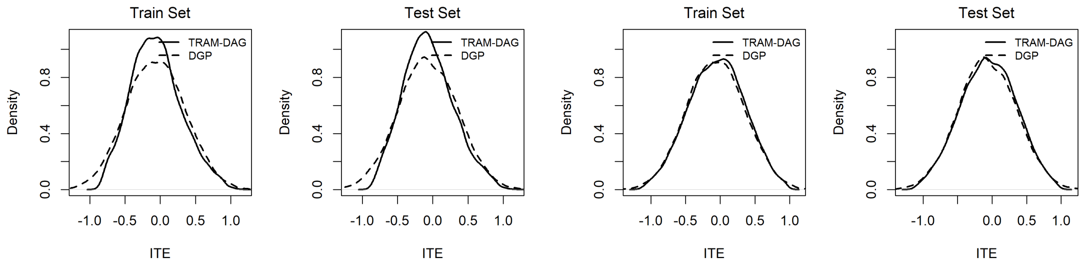


Figure 3.34: Densities of estimated ITEs compared to the true ITEs in the training and test datasets for scenario (3), without direct treatment effect but including interaction effects. Left: Observational; right: RCT setting.

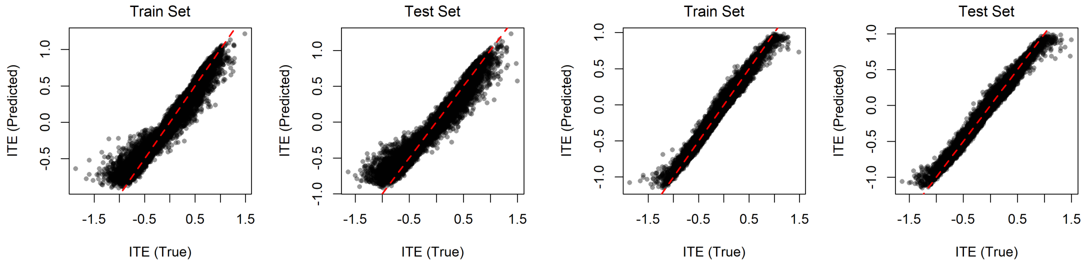


Figure 3.35: Scatterplots of estimated ITEs compared to the true ITEs in the training and test datasets for scenario (3), without direct treatment effect but including interaction effects. Left: Observational; right: RCT setting.

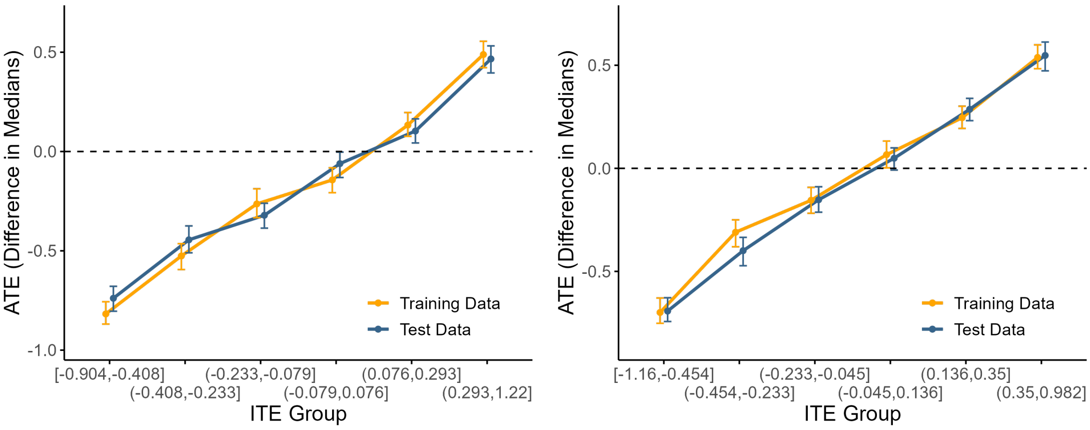


Figure 3.36: ITE-ATE plot for scenario (3), without direct treatment effect but including interaction effects. Individuals are grouped into bins according to the estimated ITE and in each bin the ATE is calculated as the difference in medians of the observed outcomes under the treatments. 95% bootstrap confidence intervals indicate the uncertainty. Left: Observational; right: RCT setting.

Chapter 4

Discussion and Outlook

Discuss all of the following subjects also include literature and reasoning and explanation (maybe also move a part to methods) (include basically anything that we encountered): ghost interactions (literature from hoogland paper (2021?) and non-collapsibility from susanne and torsten). Overfitting of certain models in differenct settings. We need enough true heterogeneity so that the model can actually detect something, complex models might just overfit else. And the relevant variables observed, else models seem to have problems to allocate the observed heterogeneity. complex models tend to predict too much heterogeneity (if there is no heterogeneity), however in the case where an interaction variable was not observed, also the complex model predicted too narrow heterogeneity (see tuned-rf scenario 2 unobserved). Talk about S learner and T learner and the difference in performance or things we have to consider? Tram dag s learner seems to also work well when fully observed. We used QTE for last experiment, but Expected potential outcomes would also be possible with sampling or numerical integration (lucas mentioned that.) ITE based on Expected is certainly more generally known etc. but in certain applictions /problems QTE might be a better choice. Because of computational simplicity we used QTE.

Check all of the following again when including the final Experiment results in section 3 (here written just from memory):

4.1 Experiment 1: TRAM-DAG (simulation study)

The tram dag can accurately estimate the causal dependencies with interpretable coefficients.

The results demonstrate that the TRAM-DAG model is able to learn the true parameters and shifts from the data, and subsequently be used as generative model to predict interventions and counterfactuals.

4.2 Experiment 2: ITE on International Stroke Trial (IST)

We observed similar results as reported by Chen *et al.* (2025) across all three models: the T-learner logistic regression, T-learner tuned random forest and the S-learner TRAM-DAG. The logistic model shows moderate discrimination in the training set, which did not generalize to the test set, as illustrated in the ITE-ATE plot in Figure 3.7. The tuned random forest model showed a stronger discrimination in the training set, but also failed to generalize to the test set (Figure 3.8). In contrast, the TRAM-DAG S-learner estimates less heterogeneity as the other two models, as shown in the density plot in Figure 3.9, resulting in a weak discrimination in both the training and test set. For all three models, the confidence intervals for the ITE-ATEs plots in the test set include the zero-line, suggesting no significant effect in any of the estimated ITE subgroup. Poor calibration does not appear to cause the limited ITE performance, as shown in the Appendix 6.6, Figures (6.1 - 6.3). However, since the ground truth is unknown, it remains unclear whether the models fail to capture true treatment effect heterogeneity, or if, for example, the underlying heterogeneity is too small, or influenced by unobserved effect modifiers. We explore this further in Experiment 3 (ITE Simulation Study).

4.3 Experiment 3: ITE model robustness under RCT conditions (simulation study)

In scenario 1, where treatment effect heterogeneity was large and all covariates were observed, the T-learner logistic regression accurately estimated the ITE. The observed ATE, conditional on the respective ITE subgroup, was well calibrated in both, the training and test dataset, as shown in the ITE-ATE plot in Figure 3.11. This is as expected, since the data was generated with the same model class (logistic regression) and applying logistic regression as T-learner for ITE estimation can accurately capture the interaction effects. The tuned random forest model also performed well. As illustrated in Figure 3.12, choosing a different model class as in the DGP, may lead to worse prediction accuracy in terms of $\mathrm{P}(Y = 1 \mid X, T)$ and ITE. This difference between the two models arises because the logistic regression model has only a small number of parameters, and with sufficient data, these parameters can converge to their true values that were used in the logistic DGP, allowing near-perfect recovery of the true probabilities and thus ITEs. In contrast, the non-parametric random forest must infer the underlying probabilities from the observed binary outcomes (0 or 1), which are themselves realizations of a Bernoulli process. This introduces inherent noise, making it harder for the model to estimate the true risk accurately - even with large sample sizes. Nonetheless, the tuned random forest also captured the general trend of the ITEs, as reflected in the ITE-ATE plot. Both models were able to capture treatment effect heterogeneity well under full observability of covariates.

In scenario 2, where the treatment effect heterogeneity remained large but one important interaction covariate (X_1) was not observed, prediction accuracy decreased for both models and the estimated heterogeneity in terms of the ITE was smaller than the true heterogeneity. Although not all heterogeneity could be recovered, the T-learner logistic regression still estimated the ITEs in the correct direction. As shown in Figure 3.14, the confidence intervals for the ATE per ITE subgroup covered the calibration line. This indicates that individuals estimated to have a smaller ITE indeed experienced worse outcomes under treatment, compared to untreated individuals in the same subgroup. Although a considerable number of individuals had a true ITE that was positive, the T-learner logistic regression did not predict a single positive ITE. This shows that the missing covariate X_1 prevents that we can detect the individuals that would actually benefit from the treatment. In contrast, the T-learner tuned random forest estimated larger treatment effect heterogeneity than the logistic model, but still could not accurately estimate the ITE and also failed to detect patients that would benefit from the treatment. The ITE-ATE plot in Figure 3.15 illustrates that the model discriminates too strongly in the training set, and does not generalize well to the test set.

In scenario 3, where the true treatment effect heterogeneity was small and all covariates were observed, the T-learner logistic regression estimated a larger heterogeneity than the truth. In the ITE-ATE plot in Figure 3.17, the confidence intervals of all ITE subgroups overlap and include the zero-line, indicating the treatment effect is not significantly different from zero. This matches expectations given the small true effect sizes. However, the T-learner tuned random forest model wrongly estimated an even larger larger treatment effect heterogeneity than the logistic regression model. As shown in Figure 3.18, the model exhibited strong discrimination in the training set but did not replicate this pattern in the test set, where regardless of the estimated ITE, the observed outcomes are similar.

Tuning more flexible models like random forests using cross-validation improved the generalization to a test set, but led to poor calibration in terms of predicted probability vs. empirically observed outcomes in the training set. An illustrative case is shown in Appendix 6.8 for the T-learner tuned random forest in scenario 3 (with weak effects), where calibration was poor in the training set but aligned well with the identity line in the test set. We observed this pattern whenever in the ITE-ATE plot the results from the training set did not generalize to the test set. This highlights the importance of evaluating models on an independent test set, when tuning a model to prevent overfitting. Although, evaluation on a test set should be done in any case.

In this experiment we showed that even though causal ML models for ITE estimation can be well calibrated in terms of prediction accuracy $\mathrm{P}(Y = 1 \mid X, T)$, they can still fail to estimate the ITE accurately under less favorable scenarios. In the case of full observability of covariates, but low interaction effects, models may estimate too high heterogeneity, which is not present in the data. However, this may become visible in the ITE-ATE plot on the test set, which can reveal that the apparent heterogeneity does not generalize. A more serious challenge arises when crucial effect-modifying variables are unobserved: in such cases, only a part of the heterogeneity can potentially be captured. Although the ITE estimates may still reflect the correct direction (i.e., be unbiased), they may fail to identify individuals who would actually benefit from treatment. Critically, this underestimation of heterogeneity is not apparent in ITE-ATE plots, making it difficult to detect in practice. Whether poor ITE performance is due to truly weak heterogeneity or to unobserved variables remains an important and open problem.

4.4 Experiment 4: ITE estimation with TRAM-DAGs (simulation study)

We analyzed ITE estimation under an observational setting (confounded) and under an RCT setting (randomized treatment allocation) in three different scenarios - direct and interaciton treatment effect, only direct but no interaction effect, and no direct but with interaction effect. We noticed that in the first scenario with

What might be surprising is that in scenario 1 where we dont have explicitly included interaction terms in the data generating process, there is still some heterogeneity in the treatment effect (as shown in figure XX). One might expect that the ITE is constant across all individuals in such a case. However since we used a non linear transformatino function as intercept in the data generating process (as would likely be the case in a real world setting), the treatment effect is not constant across all individuals (that is the ATE). When a linear transformation function would be applied (as for example a linear regression is specified, where the latent noise distribution would be the standard normal and the transformation function would be linear) then the noise term cancels out when calculating the ITE, leading to a constant ITE when no interactions are present: ITE $= \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] = (\beta_0 + \beta_t 1 + \beta_x X + \epsilon) - (\beta_0 + \beta_t 0 + \beta_x X + \epsilon) = \beta_t$.

In a model with nonlinear transformation, as in this experiment, the noise term does not cancel out anymore leading to different ITEs for patients with different characteristics.

$$\mathrm{ITE} = \mathrm{E}[Y(1) - Y(0)] = \mathrm{E}[h^{-1}(Z + \beta_t 1 + \beta_x X)] - \mathrm{E}[h^{-1}(Z + \beta_t 0 + \beta_x X)] \tag{4.1}$$

where h is the nonlinear transformation function, Z is the latent noise term, β_t is the direct treatment effect and β_x are the coefficients of the covariates. The state of the covariates X alters the position on the transformation function and thereby affects the difference between the two terms. If the transformation was fixed to be linear, the difference would be constant independent of the state of the covariates X. (This also has to do with non-collapsibility as discussed by susanne and torsten , also check Beates Mail 21.06.2025, and chatgpt discussion)

(Hoogland *et al.*, 2021) chapter 4.1 well described this phenomenon of non-additivity leaving the log-odds scale.

Chapter 5

Conclusions

The poor performance in the IST dataset was likely due to true weak heterogeneity or due to unobserved variables. We come to this conclusion because of our simulations in experiment 3 that revealed these possible problems.

We showed how TRAM-DAGS can be applied do estimate the causal relationships in a given fully observed DAG. We pointed out the importance of individualized treatment effects, for example in personnalized medicine or targeted marketing. Calibration of causal ML models is key to achieve an accurate ITE estimation. Also the trade off between complexity and generalizability becomes more important in this application compared to sole predictive modelling. We pointed out potential pitfalls that can emerge in real world settings and should be paid attention towards. These can be for example too little heterogeneity or general poor effect of the treatment, or the fact that there could be unobserved effect modifiers (treatment-covariate interactions). In terms of effect modifiers, methods in literature have already been proposed such as instrumental variables (IV) or Negative Controls (?) where additional variables in a special dependency to the treatment and exposure are used to adjust for unobserved variables (confounders or effect modifiers?). However, it strongly depends on the setting and it is not guaranteed that there exist such supporting variables. We claim that if we know the structure of the DAG, with TRAM-DAGs we can estimate the ITE regardless if we have a RCT or observational data. The only requirement is that the DAG is correct and fully observed, i.e. no unobserved confounders or effect modifiers exist. And since the average treatment effect (ATE) is the average of the individual treatment effects (ITE), we can also estimate the ATE from the ITEs. This implies that running an expensive RCT is not necessary if we have a good observational dataset and know the DAG structure. Our last experiment supports this claim. We used the medians of the potential outcomes to calculate the ITE, however, if the ITE was calculated based on the expected values, it would be directly comparable to the ATE from the RCT in terms of the difference in means, which might be a more classical measure.

Bibliography

Breiman, L. (2001). Random forests. *Mach. Learn.*, **45**, 5–32. 20

Calster, B. V., van Smeden, M., Cock, B. D., and Steyerberg, E. W. (2020). Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical Methods in Medical Research*, **29**, 3166–3178. 16

Chen, H., Aebersold, H., Puhan, M. A., and Serra-Burriel, M. (2025). Causal machine learning methods for estimating personalised treatment effects – insights on validity from two large trials. 4, 18, 19, 47

Curth, A., Peck, R. W., McKinney, E., Weatherall, J., and van der Schaar, M. (2024). Using machine learning to individualize treatment effect estimation: Challenges and opportunities. *Clinical Pharmacology & Therapeutics*, **115**, 710–719. 14

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, **95**, 407–424. 2

Frauen, D. and Feuerriegel, S. (2023). Estimating individual treatment effects under unobserved confounding using binary instruments. Accepted at ICLR 2023. 16

Freedman, B. (1987). Equipoise and the ethics of clinical research. *The New England journal of medicine*, **317**, 141–145. 1

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22. 20

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 1321–1330. PMLR. 15

Gutierrez, P. and Gérardy, J.-Y. (2017). Causal inference and uplift modelling: A review of the literature. In Hardgrove, C., Dorard, L., Thompson, K., and Douetteau, F., editors, *Proceedings of The 3rd International Conference on Predictive Applications and APIs*, volume 67 of *Proceedings of Machine Learning Research*, 1–13. PMLR. 2

Hariton, E. and Locascio, J. J. (2018). Randomised controlled trials - the gold standard for effectiveness research. *BJOG: An International Journal of Obstetrics & Gynaecology*, **125**, 1716 – 1716. 1

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep iv: a flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 1414–1423. JMLR.org. 16

Herzog, L., Kook, L., Götschi, A., Petermann, K., Hänsel, M., Hamann, J., Dürr, O., Wegener, S., and Sick, B. (2023). Deep transformation models for functional outcome prediction after acute ischemic stroke. *Biometrical Journal*, **65**, 2100379. 10

Hoogland, J., IntHout, J., Belias, M., Rovers, M. M., Riley, R. D., E. Harrell Jr, F., Moons, K. G. M., Debray, T. P. A., and Reitsma, J. B. (2021). A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Statistics in Medicine*, **40**, 5961–5981. 19, 49

Hothorn, T., Kneib, T., and Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **76**, 3–27. 4, 7

Hothorn, T., Möst, L., and Bühlmann, P. (2018). Most likely transformations. *Scandinavian Journal of Statistics*, **45**, 110–134. 10, 25

International Stroke Trial Collaborative Group (1997). The international stroke trial (ist): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19,435 patients with acute ischaemic stroke. *The Lancet*, **349**, 1569–1581. 18

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. 10, 17

Nichols, A. (2007). Causal inference with observational data. *The Stata Journal*, **7**, 507 – 541. 1, 16

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**, 669–688. 12

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys*, **3**, 96 – 146. 1

Pearl, J. (2009b). *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition. 2, 3, 13

Poinsot, A., Leite, A., Chesneau, N., Sébag, M., and Schoenauer, M. (2024). Learning structural causal models through deep generative models: methods, guarantees, and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24. 4

Reisach, A. G., Seiler, C., and Weichwald, S. (2021). Beware of the simulated dag! causal discovery benchmarks may be easy to game. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21. Curran Associates Inc., Red Hook, NY, USA. 11

Riley, R. D., Snell, K. I., Martin, G. P., Whittle, R., Archer, L., Sperrin, M., and Collins, G. S. (2021). Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *Journal of Clinical Epidemiology*, **132**, 88–96. 16

Sandercock, P. A., Niewada, M., Członkowska, A., and the International Stroke Trial Collaborative Group (2011). The international stroke trial database. *Trials*, **12**, 101. 18

Sick, B. and Dürr, O. (2025). Interpretable neural causal models with tram-dags. Accepted at the CLeaR 2025 Conference. 1, 4

Sick, B., Hathorn, T., and Dürr, O. (2021). Deep transformation models: Tackling complex regression problems with neural network based transformation models. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2476–2481. 4, 7, 9

Zhao, Z. and Harinen, T. (2020). Uplift modeling for multiple treatments with cost optimization. 2

Chapter 6

Appendix

To do:

- include results of ITE estimation with not tuned RF for scenario 1 (fully observed,strong effects) to show the importance that models should be well calibrated (in this case tuned) to yield good results

- inlcude Example of ITE simulation with tram dags with complex shift: $theta + LS(X2) + CS(T, X1)$ to show how tram dags can model interactions. Maybe just refer in the methods for the CS of tram dags, or alternatively in the ITE with tram dags (where i used CI, but to show that CS is also possible to allow for specific interactions)

Complex shift (Interaction example) to show what is also possible:

Here I just want to make a short input from another example. So there the true model was that of a logistic regression with the binary outcome Y and 3 predictors. The binary treatment T and the two continuous predictors X1 and X2. There was also an interaction effect assumed between treatment and X1. So this basically means that the effect of X1 on the outcome is different for the two treatment groups. And here we can show that our TRAM-DAG specified by a complex shift of T and X1 can also capture this interaction effect quite well.

6.1 Negative Log Likelihood

6.1.1 Continuous Outcome

For a continuous outcome Y the CDF is given by:

$$F_{Y|\mathbf{X}=\mathbf{x}}(y) = F_Z(h(s(y) \mid \mathbf{x})) \tag{6.1}$$

where in our case F_Z is the cumulative distribution function of the standard logistic distribution

$$F_Z(z) = \frac{1}{1+e^{-z}}, \quad z \in \mathbb{R} \tag{6.2}$$

and h is the conditional transformation function that maps the scaled outcome $s(y)$ to the latent scale Z (log-odds).

The outcome y has to be scaled onto the range $[0, 1]$, because the Bernstein polynomial is bounded:

$$s(y) = \frac{y - \min(y)}{\max(y) - \min(y)} \tag{6.3}$$

This scaling also has to be considered when taking the derivative to get the PDF with the change of variables formula:

$$f_{Y|\mathbf{X}=\mathbf{x}}(y) = f_Z(h(s(y) \mid \mathbf{x})) \cdot h'(s(y) \mid \mathbf{x}) \cdot s'(y) \tag{6.4}$$

Where f_Z is the PDF of the standard logistic distribution:

$$f_Z(z) = \frac{e^z}{(1+e^z)^2}, \quad z \in \mathbb{R} \tag{6.5}$$

Finally, the NLL-contributions are then given by the negative log-densities evaluated at the observations.

$$\text{NLL} = -\log(f_{Y|\mathbf{X}=\mathbf{x}}(y)) \tag{6.6}$$

The full formula is given by

$$\begin{aligned}\text{NLL} = -\log f_{Y|\mathbf{X}=\mathbf{x}}(y) &= -h(s(y) \mid \mathbf{x}) - 2\log(1+\exp(-h(s(y) \mid \mathbf{x}))) \\ &\quad + \log h'(s(y) \mid \mathbf{x}) - \log(\max(y) - \min(y))\end{aligned} \tag{6.7}$$

6.1.2 Discrete Outcome

The for a discrete outcome (binary, ordinal, categoric) with categories y_k, $k = 1, \dots, K$, the CDF is given by:

$$F(Y_k \mid \mathbf{X}) = F_Z(h(y_k \mid \mathbf{x})) \tag{6.8}$$

The likelihood contributions are then given by

$$l_i(y_k \mid \mathbf{x}) = f_{Y_k|\mathbf{X}=\mathbf{x}}(y_k) = \begin{cases} F_Z(h(y_k \mid \mathbf{x})) & k = 1 \\ F_Z(h(y_k \mid \mathbf{x})) - F_Z(h(y_{k-1} \mid \mathbf{x})) & k = 2, \dots, K-1 \\ 1 - F_Z(h(y_{k-1} \mid \mathbf{x})) & k = K \end{cases} \tag{6.9}$$

from which the NLL-contributions are derived

$$\text{NLL} = -\log(f_{Y_k|\mathbf{X}=\mathbf{x}}(y) \tag{6.10}$$

6.2 Interpretation of Linear Coefficients

The transformation model framework allows for interpretation of the coefficients in the linear shift. Consider the conditional transformation function:

$$F_{X_2|X_1}(x_2) = \text{expit}(h(x_2) + \beta_{12}x_1), \tag{6.11}$$

where $h(x_2)$ is a smooth, monotonic transformation (e.g., a Bernstein polynomial), and β_{12} is a linear coefficient encoding the effect of X_1 on X_2.

Taking the logit (inverse expit) yields:

$$\log\left(\frac{F_{X_2|X_1}(x_2)}{1 - F_{X_2|X_1}(x_2)}\right) = h(x_2) + \beta_{12}x_1. \tag{6.12}$$

This linear additive structure allows the interpretation of β_{12}. The odds ratio when increasing x_1 by one unit is:

$$\text{OR}_{x_1 \to x_1+1} = \frac{\exp(h(x_2) + \beta_{12}(x_1+1))}{\exp(h(x_2) + \beta_{12}x_1)} = \exp(\beta_{12}). \tag{6.13}$$

Interpretation: The quantity $\exp(\beta_{12})$ represents the **multiplicative change in odds** for $X_2 \leq x_2$ when increasing X_1 by one unit, holding all else constant.

6.3 Encoding of discrete variables

In the TRAM-DAG a variable X_i can act as a predictor variable for a child node, or as an outcome (child node) that depends on some parent nodes. When X_i is acting as an outcome, the distribution of the variable X_i represented by the transformation function h which estimates a cut-point for each variable. So different form of intercept h_i is used compared to a continuous outcome variable.

If a discrete variable X_i with K categories is used as a predictor variable, it should be dummy encoded. This is done by creating $K - 1$ binary variables, where each variable indicates whether the observation belongs to this specific category/level or not. The first category/level is used as the reference and is not explicitly included in the model.

Example: for an ordinal variable X_i with three levels (1, 2 3), we create two binary variables:

- $X_{i,1}$: 1 if $X_i = 2$, 0 otherwise
- $X_{i,2}$: 1 if $X_i = 3$, 0 otherwise

Assume a continuous outcome Y that depends on the ordinal variable X with 3 levels, the CDF for Y is given by: $F(Y \mid X = 1) = F_Z(h_I(y) + x_1\beta_1 + x_2\beta_2)$

For $X = 1$, the reference level, the CDF simplifies to: $F(Y \mid X = 1) = F_Z(h_I(y))$

For $X = 2$, the CDF becomes: $F(Y \mid X = 1) = F_Z(h_I(y) + \beta_1)$

For $X = 3$, the CDF becomes: $F(Y \mid X = 1) = F_Z(h_I(y) + \beta_2)$

The coefficients β_1 and β_2 can be interpreted as the additive shift in the latent scale $h_I(y)$ when moving from the reference level (1) to levels 2 and 3, respectively.

6.4 Scaling of continuous variables

Neural networks work best when the input variables are standardized. A linear, monotonic and invertible transformation of a predictor variable changes the interpretation of the coefficient. Scaling a predictor variable X as $X_{\text{std}} = (X - mean(X))/sd(X)$ will imply that the coefficient $\tilde{\beta}$ is interpreted as the change in log-odds for a one standard deviation increase in the predictor variable or equivalently, for a one unit increase in the standardized predictor. This is different from the interpretation of the coefficient β in the original scale, which represents the change in log-odds for a one unit increase in the predictor variable.

In contrast, the standardization of the outcome variable has no effect on the interpretation (because the scale invariance of the log-odds). Consider, we standardize the outcome Y as follows:

$$Y_{\text{std}} = \frac{Y - \mu_Y}{\sigma_Y}$$

This transformation is linear, monotonic, and invertible:

$$Y = Y_{\text{std}} \cdot \sigma_Y + \mu_Y$$

Therefore, for any threshold y, we have the equivalence:

$$P(Y < y \mid X) = P\left(Y_{\text{std}} < \frac{y - \mu_Y}{\sigma_Y} \mid X\right)$$

This means that the probability is the identical when evaluating the same quantile in the standardized outcome as in the raw outcome. Furthermore, the interpretation of coefficients in a continuous outcome logistic regression remains unchanged. In particular, the log-odds ratio:

$$\log\left(\frac{P(Y<y\mid X+1)}{1-P(Y<y\mid X+1)}\right)-\log\left(\frac{P(Y<y\mid X)}{1-P(Y<y\mid X)}\right)$$

is equal to:

$$\log\left(\frac{P\left(Y_{\text{std}}<\frac{y-\mu_Y}{\sigma_Y}\mid X+1\right)}{1-P\left(Y_{\text{std}}<\frac{y-\mu_Y}{\sigma_Y}\mid X+1\right)}\right)-\log\left(\frac{P\left(Y_{\text{std}}<\frac{y-\mu_Y}{\sigma_Y}\mid X\right)}{1-P\left(Y_{\text{std}}<\frac{y-\mu_Y}{\sigma_Y}\mid X\right)}\right)$$

as long as the same quantile (i.e. probability threshold) is used. Thus, the coefficient β reflects the same change in log-odds for a one-unit increase in the (standardized) predictor, regardless if the outcome is standardized or not. This property is also crucial for the evaluation of the bernstein polynomial, since the outcome has to be scaled on a range between 0 and 1.

The general formula of the transformation model is

$$P(Y<y\mid X=x)=F_z\left(h(Y)+\beta\cdot X\right)$$

but the model is fitted with standardized outcome and predictors

$$P(Y_{\text{std}}<y_{\text{std}}\mid X_{\text{std}}=x_{\text{std}})=F_z\left(\tilde{h}(Y_{\text{std}})+\tilde{\beta}\cdot X_{\text{std}}\right)$$

where $\tilde{h}$ and $\tilde{\beta}$ represent the estimated transformation function and coefficients after standardizing the outcome and predictors.

For example, if we want to know the probability $P(Y<20\mid X=3)$ with standardized variables, the model is specified as

$$P\left(\frac{Y-\mu_Y}{\sigma_Y}<\frac{20-\mu_Y}{\sigma_Y}\,\middle|\,X_{\text{std}}=\frac{3-\mu_X}{\sigma_X}\right)=F_z\left(\tilde{h}\left(\frac{20-\mu_Y}{\sigma_Y}\right)+\tilde{\beta}\cdot\frac{3-\mu_X}{\sigma_X}\right)$$

6.5 Bernstein Polynomial for Continuous Outcomes

In deep TRAMs the intercept for continuous variables is a smooth monotonically increasing function that is represented by a Bernstein polynomial of order K (here the complex intercept case where the Intercept already depends on the predictors x, however, the same principle that follows also applies for the simple intercept case):

$$h_I(y\mid \mathbf{x})=\sum_{k=0}^{K}b_k(\mathbf{x})\cdot B_k(s(y)) \tag{6.14}$$

where $B_k(s(y))$ is the Bernstein basis polynomial of order K evaluated at the scaled outcome $s(y)$:

To guarantee that the transformation $h_I(y\mid \mathbf{x})$ is monotonically increasing in y, the coefficients $b_k(\mathbf{x})$ must form a non-decreasing sequence. This is ensured via a *cumulative softmax* parameterization. Instead of learning $b_k(\mathbf{x})$ directly as the outputs of the intercept neural network, we first define unbounded parameters $\theta_k(\mathbf{x})\in\mathbb{R}$ and then compute the Bernstein parameters using the cumulative softmax transformation:

$$\tilde{b}_k(\mathbf{x})=\sum_{j=0}^{k}\frac{\exp(\theta_j(\mathbf{x}))}{\sum_{\ell=0}^{K}\exp(\theta_\ell(\mathbf{x}))},\quad \text{for } k=0,\ldots,K. \tag{6.15}$$

This transformation produces a vector $\tilde{b}_k(\mathbf{x})$ that is monotonically increasing in k, with values bounded in $[0,1]$. It ensures that:

- $\tilde{b}_0(\mathbf{x}) \leq \tilde{b}_1(\mathbf{x}) \leq \ldots \leq \tilde{b}_K(\mathbf{x})$, - The sum of increments is 1, - The transformation is smooth and differentiable.

The combination of Bernstein polynomials with cumulative softmax-transformed parameters allows flexible, smooth, and strictly monotonic transformations of continuous outcomes, which are essential properties for distribution estimation and generative sampling within the deep TRAM architecture.

6.5.1 Scaling and Extrapolation of the Bernstein Polynomial

Because the Bernstein polynomial is only defined on the range $[0, 1]$ the outcome y has to be scaled onto the same range. Furthermore, for the sole purpose of estimating the parameters of the Bernstein polynomial it would be sufficient to finish here. However, one has to be able to also evaluate $h(y \mid \mathbf{x})$ for arbitrary values of y, in particular also the ones that are outside of $(\min(y_{train}), \max(y_{train})))$. This is also crucial for sampling. Therefore we extend the Bernstein polynomial by linearly extrapolating the tails of the polynomial. We do this by constructing inside the 5% and 95% quantiles of y by the smooth Bernstein polynomial 6.14 and linearly extrapolating the outside this range using the slope of the polynomial at the boundaries. This results in a piecewise-defined function that is differentiable, monotonic, and defined for all real values of y, which is essential for evaluating the model at arbitrary outcomes or for generative sampling.

Tho formalize this, let $q_{0.05}$ and $q_{0.95}$ denote the 5% and 95% empirical quantiles of the outcome y, computed on the training data. The scaled outcome is defined as

$$s(y) = \frac{y - q_{0.05}}{q_{0.95} - q_{0.05}}. \tag{6.16}$$

This scaling maps the interval $[q_{0.05}, q_{0.95}]$ to the unit interval $[0, 1]$, which is the domain of the Bernstein basis polynomials. Let $h_I(s(y) \mid \mathbf{x})$ be the original transformation as defined in Equation (6.14). The extrapolated transformation $h^*(y \mid \mathbf{x})$ is then defined as

$$h^*(y \mid \mathbf{x}) = \begin{cases} h_I(0 \mid \mathbf{x}) + h_I'(0 \mid \mathbf{x}) \cdot (s(y) - 0), & \text{if } s(y) < 0 \\ h_I(s(y) \mid \mathbf{x}), & \text{if } 0 \leq s(y) \leq 1 \\ h_I(1 \mid \mathbf{x}) + h_I'(1 \mid \mathbf{x}) \cdot (s(y) - 1), & \text{if } s(y) > 1 \end{cases} \tag{6.17}$$

The function is thus extrapolated beyond the central range using the tangent line at the boundaries. The derivatives $h_I'(0 \mid \mathbf{x})$ and $h_I'(1 \mid \mathbf{x})$ are computed analytically from the Bernstein basis and the learned coefficients $b_k(\mathbf{x})$, and ensure continuous differentiability across the domain (see next subsection).

This construction ensures several desirable mathematical properties. First, the transformation $\tilde{h}(y \mid \mathbf{x})$ is globally defined on $\mathbb{R}$, avoiding undefined regions or discontinuities. Second, it preserves monotonicity due to the use of the cumulative softmax parameterization of the coefficients $b_k(\mathbf{x})$, which guarantees that the Bernstein polynomial is strictly increasing. Finally, the piecewise-linear extrapolation ensures the function is continuously differentiable and smooth at the junctions $s(y) = 0$ and $s(y) = 1$.

6.5.2 Analytical Derivative of the Bernstein Polynomial Transformation

To efficiently compute the gradient of the transformation $h_I(y \mid \mathbf{x})$ with respect to its inputs, we can exploit the analytical structure of the Bernstein basis polynomials. Recall the general form of the transformation:

$$h_I(y \mid \mathbf{x}) = \sum_{k=0}^{K} b_k(\mathbf{x}) \cdot B_k(s(y)), \tag{6.18}$$

where $B_k(s)$ are the Bernstein basis polynomials of order K, and $b_k(\mathbf{x})$ are predictor-dependent coefficients. For fixed $\mathbf{x}$, the derivative with respect to y is needed, for example, to evaluate the density function when h_I is used in a generative model.

Let us denote $s = s(y)$. Using the chain rule, we compute:

$$\frac{d}{dy} h_I(y \mid \mathbf{x}) = \sum_{k=0}^{K} b_k(\mathbf{x}) \cdot \frac{d}{dy} B_k(s) = \sum_{k=0}^{K} b_k(\mathbf{x}) \cdot \frac{dB_k(s)}{ds} \cdot \frac{ds}{dy}. \tag{6.19}$$

The derivative of the scaled outcome $s(y) = \frac{y - q_{0.05}}{q_{0.95} - q_{0.05}}$ is simply

$$\frac{ds}{dy} = \frac{1}{q_{0.95} - q_{0.05}}. \tag{6.20}$$

The derivative of the Bernstein basis polynomial $B_{k,K}(s)$ is known and given by:

$$\frac{d}{ds} B_{k,K}(s) = K \left[B_{k-1,K-1}(s) - B_{k,K-1}(s) \right]. \tag{6.21}$$

Therefore, the full derivative is:

$$\frac{d}{dy} h_I(y \mid \mathbf{x}) = \frac{K}{q_{0.95} - q_{0.05}} \sum_{k=0}^{K} b_k(\mathbf{x}) \left[B_{k-1,K-1}(s) - B_{k,K-1}(s) \right]. \tag{6.22}$$

This expression can be evaluated efficiently and is used both in the likelihood computation (e.g., via change-of-variables) and for constructing tail extrapolations with matching slopes.

6.6 Calibration plots: Experiment 2

Figure 6.1 - 6.3 show the calibration plots in terms of the predicted risks against the the observed proportions for the models applied in experiment 2 (International Stroke Trial (IST)).

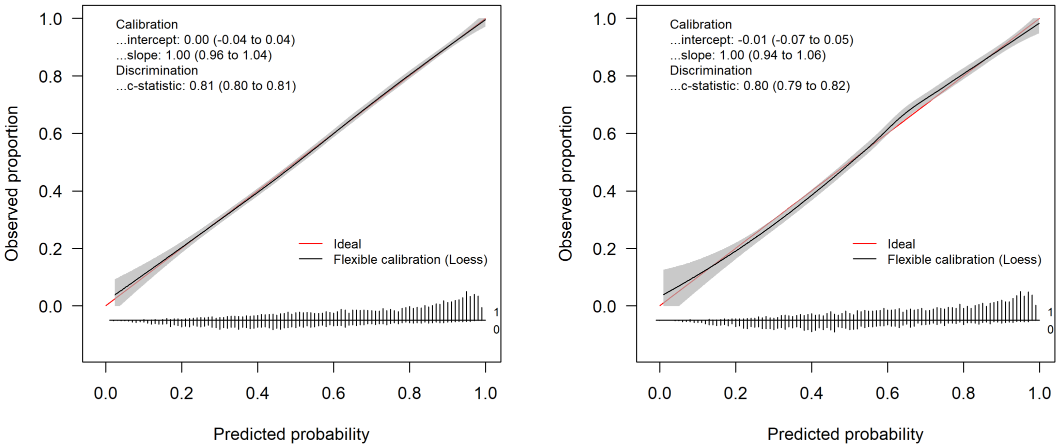


Figure 6.1: Calibration plot for the T-learner logistic regression applied on the International Stroke Trial (IST) in experiment 2. It shows the predicted risks against the the observed proportions of the event. Left: training dataset; Right: test dataset.

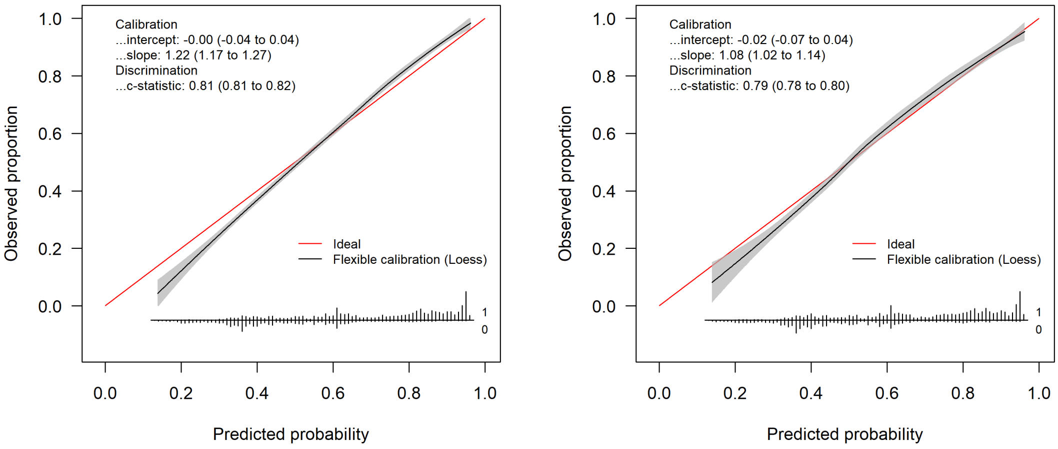


Figure 6.2: Calibration plot for the T-learner tuned random forest applied on the International Stroke Trial (IST) in experiment 2. It shows the predicted risks against the the observed proportions of the event. Left: training dataset; Right: test dataset.

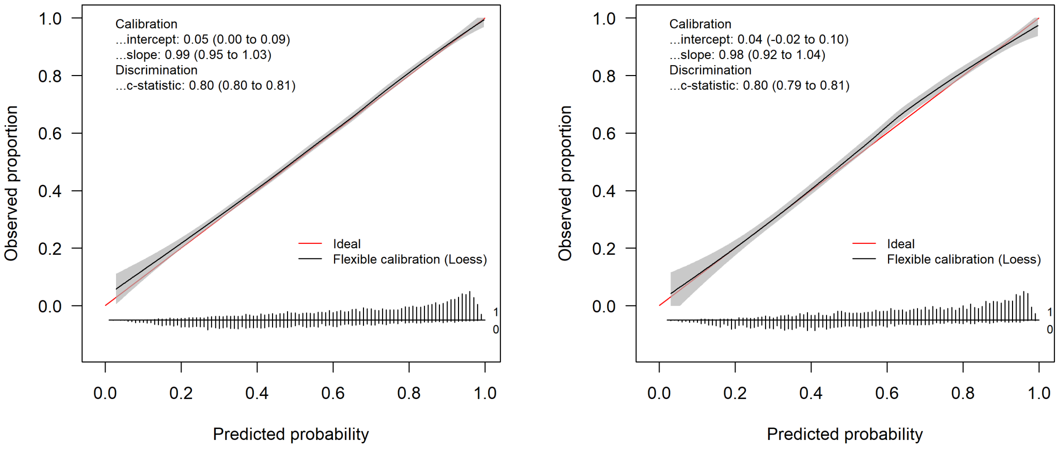


Figure 6.3: Calibration plot for the S-learner TRAM-DAG applied on the International Stroke Trial (IST) in experiment 2. It shows the predicted risks against the the observed proportions of the event. Left: training dataset; Right: test dataset.

6.7 Default Random Forest for ITE Estimation

In Section 2.2.3 we pointed out the importance of calibration of models when estimating individual treatment effects. In this section we show the results of the default random forest model without tuning for scenario (1), illustrated in Figure 6.4 where all variables are observed and there are strong treatment and interaction effects. The results are shown in Figure 6.5. In the scatterplot of true vs. predicted probabilities for $\mathrm{P}(Y_i = 1 \mid \mathbf{X_i} = \mathbf{x_i}, T_i = t_i)$ in the train set, it is visible that the model does not predict the probabilities accurately, hence is not well calibrated. This poor calibration also translates to the estimated ITEs. In comparison, the results of the tuned random forest in Figure 3.12 show that the model is better calibrated and the estimated ITEs are close to the true ITEs. This illustrates the importance of tuning models for ITE estimation, as poor calibration can lead to biased estimates of individualized treatment effects.

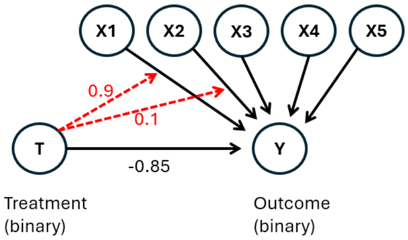


Figure 6.4: DAG for scenario (1), where all variables are observed and there are strong treatment and interaction effects. The numbers indicate the coefficients on the log-odds-scale. Red: interaction effects between treatment (T) and covariates (X_1 and X_2) on the outcome (Y).

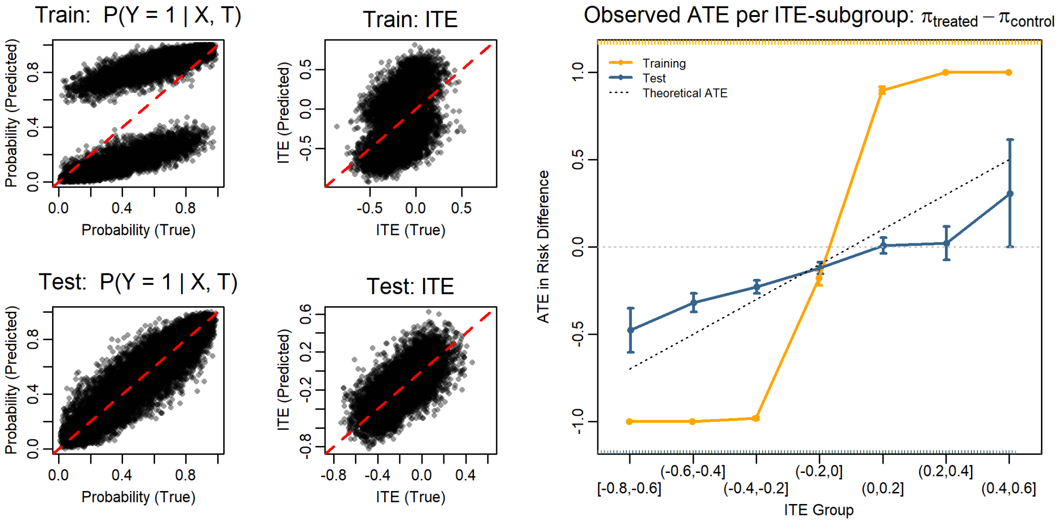


Figure 6.5: Results with the default random forest in scenario (1) when the DAG is fully observed and there are strong treatment and interaction effects. Left: true vs. predicted probabilities for $P(Y = 1 \mid X, T)$; Middle: true vs. predicted ITEs; Right: observed ATE in terms of risk difference per estimated ITE subgroup.

6.8 Calibration differences for complex model: Experiment 3

Figure 6.6 shows the calibration plots in terms of the predicted risks against the the observed proportions of the event for the T-learner tuned random forest for scenario (3) with weak direct and interaction treatment effects. This is in contrast to the prediction plots presented in Section 3.3 where we presented the true probabilities of the event $P(Y = 1 \mid X, T)$ against the predicted probabilities. It becomes apparent, that tuning the random forest model out-of-bag leads to a poor calibration on the training set, but due to better generalization it leads to a better calibration on the test set.

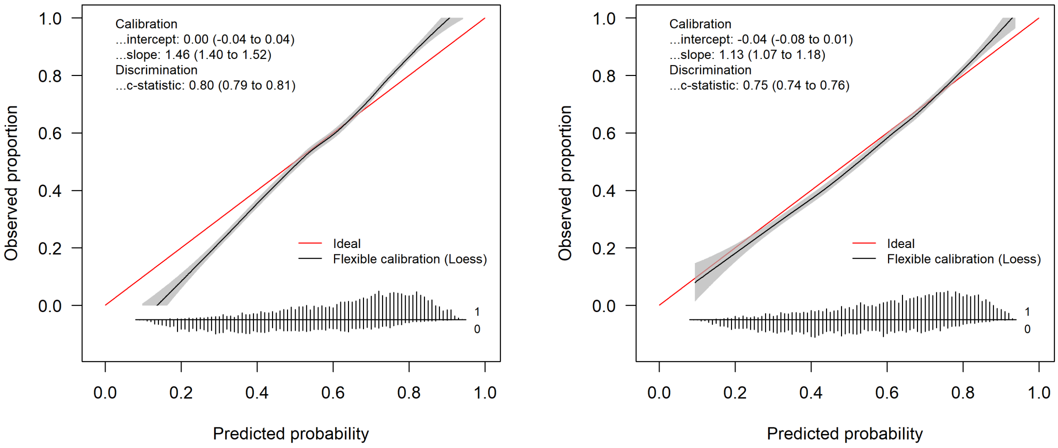


Figure 6.6: Calibration plot for the T-learner tuned random forest for scenario (3) with weak direct and interaction treatmetn effects. It shows the predicted risks against the the observed proportions of the event. Left: training dataset; Right: test dataset.