university of groningen / faculty of mathematics and natural sciences / artificial intelligence

Development and Evaluation of Causal Models With an Application to Orthopaedic Inquiries

Master's Thesis
(Computational Intelligence)

Richard Fischer (s3124908)

January 23, 2023

Internal Supervisor(s): Prof. Dr. Herbert Jaeger (Artificial Intelligence, University of Groningen)
First External Supervisor: Dr. Hamid Ghaednia (Harvard University, SORG, Boston)
Second External Supervisor: Dr. Soheil Ashkani Esfahani (Harvard University, FARIL, Boston)

Department of Artificial Intelligence, Bernoulli Institute
University of Groningen, The Netherlands
Harvard University, USA

Contents

Acknowledgments

I’m grateful for the continuous support of Prof. Dr. Herbert Jaeger, who enabled me to start my journey to the USA and supported me throughout. Furthermore, I could not have undertaken this journey without the help and guidance of Dr. Hamid Ghaednia and Dr. Soheil Ashkani-Esfahani, my supervisors at the Harvard Medical School. Special thanks to the Editors of the journal ’Clinical Orthopaedics and Related Research’ (CORR), namely Dr. Leopold, Dr. Porcher, and Dr. Wongworawat, who gave helpful feedback to the medical application of causal inference. Also, special thanks to Brady Neal, who was a PhD candidate at the Mila - Quebec AI Institute. His easily understandable video material and reader were my first stepping stones in causal inference.

Abstract

Several influential leaders in the field of Artificial Intelligence argue that something is missing in the current Deep Learning approach: causal understanding. In the field of causal reasoning, models that quantify the causal effect of some variable on another are developed. This thesis offers an overview of the field, a comparison of several estimators, including an extension to the X-Learner coined X-Learner++ and an ensemble consisting of three estimators. The comparison is done on the semi-synthetic IHDP dataset. Furthermore, a medical dataset to investigate the effect of prophylactic treatment on Venous Thromboembolism (VTE) incidence is analyzed and compared to a peer-reviewed meta-analysis. The best-performing estimator in the IHDP dataset is the Augmented Inverse Probability Weighting (AIPW) estimator and the X-Learner++ with an error of -0.09. In the medical data set, the ensemble approach worked best when compared to the meta-analysis with an error of 0.51%. The thesis introduces the causal reasoning methodology to the realm of orthopaedics, and establishes trust by successfully emulating an existing meta-analysis. Furthermore, we establish that the estimators, and the developed error correction for the X-Learner, work well on the semi-synthetic dataset.

1 Introduction

The Deep Learning (DL) revolution has had an immense impact on the field of Artificial Intelligence (AI), particularly in pattern recognition. It has enabled self-driving cars [1], sophisticated language models [2, 3], and a new age of digital art [4, 5]. The progress of learning-based AI led Rich Sutton to learn and publish 'the Bitter Lesson' [6], in which he argues that 'building in [AI systems] how we think we think does not work in the long run'. In Suttons opinion, rule-based AI will never exceed learning-based AI in the long term due to exponentially falling costs per unit compute, as *Moore's law* prescribes.

Yet, several influential leaders in the field of AI are critical of the success and are suspecting a missing element in the approach. Joshua Tenenbaum states that pattern recognition is important in developing intelligent systems, but that there is more to an intelligent system than simply recognizing patterns [7]. Furthermore, Judea Pearl thinks that the notion underlying deep learning that everything can be learned from data is insufficient. Additionally, Pearl claims, causal reasoning and understanding is required to bridge the gap between pattern recognition and intelligence [7, 8]. Schölkopf and Kügelgen (2022) argue that the causal view is a relevant building block to address open problems in the field, specifically concerning robustness and generalizability beyond the training distribution [9].

Current state-of-the-art models rely on statistical learning theory, and the assumption that the test data is *independent and identically distributed (i.i.d.)* compared to the training data. If this assumption is violated, learning guarantees cease to hold [9]. The science of causality attempts to increase the robustness of violations of this assumption by intervening on the training data, which is equivalent to a distribution shift.

Interventions are common in many fields, such as agriculture, economics, epidemiology, and medicine in general. Therefore, much work on the evaluation of an intervention, or causal reasoning, has been done in these fields [10, 11, 12, 13, 14]. The gold standard to answer interventional questions such as 'How effective is the new drug T in treating disease Y?' is the randomized control trial (RCT). In theory, the random assignment of a population to a treatment and non-treatment group removes all potential confounding, and, thereby, all third variables that may explain the effect on the disease. Yet, conducting RCTs is expensive, time intensive, and can be unethical to conduct. Causal reasoning offers a methodology to emulate such RCTs from observational data.

Many estimators to measure the effect of an intervention have been developed [11, 15, 16, 17, 18] but the result can hardly be verified. This is because causal reasoning strives to compare the same unit in two inherently different, and mutually exclusive situations: one in which the unit received the treatment to one in which the unit did not receive the treatment. A ground truth to calculate classical metrics like accuracy or mean-squared-error is hence not available. This was coined the "*fundamental problem of causal inference*" [19]. In this Master project, I suggest an error correction to the X-Learner [16] coined the X-Learner++. This extension tests the induced error after an imputation step by using the imputed values to predict the observed values. Furthermore, an ensemble approach that combines multiple estimators is evaluated.

Due to the absence of a ground truth, semi-synthetic datasets with simulated outcomes are popular tools in causal reasoning to evaluate estimators. To showcase the effectiveness of the X-Learner++, the ensemble, and three singleton estimators, the effect of an intervention aimed at improving the cognitive function of infants is evaluated [20]. This dataset has simulated outcomes, which makes a comparison to the ground truth possible.

When applying causal reasoning estimators to real-world medical datasets, such luxury is not available. Yet, it is possible to compare the result of an estimator to an already published RCT, which the model should emulate. Therefore, this thesis emulates an RCT on the efficacy of prophylactic treatment for Venous Thromboembolism (VTE), the formation of potentially lethal blood clots, after an isolated ankle fracture. The medical problem statement is explained in Section 3.2.2 and in [21]. The results are then compared to an already existing RCT to establish trust in the performance of the estimators. Then, the treatment effect for two sub-populations is calculated to investigate the effect of Statins, a widely consumed drug in cardiovascular patients, on the efficacy of VTE prophylaxis. This effort resembles one of the first attempts to apply causal reasoning in the realm of orthopaedics.

This work was created at the SORG and FARIL research collaborative at the Harvard Medical School in Boston. I stayed there for eight months and was supervised by Dr. Hamid Ghaednia and Dr. Soheil Ashkani-Esfahani. The project was partly fundeded by the Marco Polo grant from the University of Groningen. Parts of this work are currently under review at the Journal of Orthopaedic Research.

1.1 Research Questions

To summarize, this thesis focuses on the following problems:

Theoretical Research Questions

Q1. Does the error correction extension to the X-Learner reduce the error in the estimation of causal effects?

Q2. Does an ensemble of multiple estimators have less error than any of the estimators has on its own?

Medical Research Questions

Q3. Does the causal effect of VTE prophylaxis on VTE incidence coincide with the results of a meta-analysis of RCTs on the same topic?

Q4. What is the effect of Statins on the efficacy of VTE prophylaxis on the VTE incidence?

1.2 Thesis Outline

Section 2 explains the science of causality in depth with a focus on causal reasoning. Then, Section 3 introduces the estimators under investigation in this thesis, including the extension to the X-Learner. Section 3 also contains a description of the datasets on which causal effects are estimated, the medical problem statement, and a variety of applied methods. Section 4 contains specific guidelines on how the experiments are set up. Section 5 contains an overview of the results on the different datasets. Thereafter, the results are discussed in Section 6. Last, a conclusion is drawn, the contributions of this thesis are summarized, and paths for future research are outlined in Section 7.

2 Background

The following Section explains the dissimilarities between conventional statistics and causality, including an overview of the prevalent causal frameworks and independence assumptions necessary for interventions. The explanation provided follows [9].

2.1 From Statistical to Causal Models

The success of machine learning (ML) can be accounted for by the unprecedented availability of large data sets, the flexibility of modern ML models, and the computing power available today. The combination of these three factors enables the ML models to approximate complex functions by tuning the many available parameters on many examples in a reasonable amount of time. Furthermore, current approaches rely on the assumption that the data encountered in training and deployment is independent and identically distributed (i.i.d.), which is crucial for performance.

When the i.i.d. assumption is violated, general statistical learning guarantees cease to hold. For instance, vision systems can be grossly misled by adversarial attacks [22]. These attacks, invisible to the human eye, may lead a street sign detection model to predict, with high accuracy, the wrong street sign, or even the wrong speed. Furthermore, another object detection model can be misled by presenting the target object, which is normally recognized accurately, in an unfamiliar environment. These examples highlight the need to construct systems that do not solely rely on statistical dependencies. Causality provides a framework for distributional shifts, and, hence, the means to reason outside of the known.

The old statistical mantra that correlation is not causation must be addressed when discussing causality. A prominent example that illustrates this mantra is the positive correlation between Nobel prizes and chocolate consumption per capita [23]. These so-called spurious correlations can not reasonably be assumed to be causing one another. As the example suggests, the correlation seems insufficient to infer causation. But what does "causation" mean? Is there a connection between correlation and causation? And what is sufficient to infer causality?

In this Master project, a definition of causality in terms of manipulability and intervention is adopted.

Definition 3.1 (Causal Effect). *A random variable (RV) X has a causal effect on a random variable Y if there exist at least two values of X, x and x' where $x \neq x'$ s.t. the distribution of Y after intervening on X and setting it to x differs from the distribution of Y after setting X to x'.*

In other words, if the value of some RV X is changed from value x to x', and there is no change in the RV Y, X does not have a causal effect on Y. For instance, when a new drug is tested in human trials, it is expected that intervening on one group by giving them the new drug ($X = x$) will reduce the severity of some disease, compared to the group that has not received the drug ($X = x'$). While this example assumes two different groups, causal inference attempts to emulate such conditions based on observational data.

Machine learning models learn the response in the RV Y from individuals with similar levels of confounding and different levels of X (e.g. x and x'). These models are used to impute how an individual's RV Y would have changed if the RV X took another value. Specifically, every individual that has $X = x$ and some value of Y is compared to the same individual after an intervention that sets $X = x'$. This intervention is simulated by setting the RV X to x' manually. Then, previously trained models that learned the relationship between individuals with $X = x'$ and Y impute the current individual's response to Y, if that individual had $X = x'$, instead of the observed $X = x$.

To illustrate this concept further, assume two different genes X_A and X_B, which correlate with equal magnitude with the phenotype Y. When intervening on both genes by knocking them out, the phenotype only changes after the intervention on gene X_A, not after knocking out gene X_B. Therefore, X_A has a causal effect on Y, while the correlation between X_B and Y stems from a different (confounded) causal structure. These causal relationships are usually represented using *causal graphs*, which are directed acyclic graphs (DAGs) whose arrows indicate a direct causal effect.

While X_B is not a direct cause of Y, the correlation is a by-product of the causal structure determining both, X_B and Y. This connection between correlation and causation was termed the Common Cause Principle by Reichenbach [24]:

Principle 3.1 (Common Cause). *If two RVs X and Y are statistically dependent ($X \not\perp\!\!\!\perp Y$), then there exists an RV Z which causally influences both of them and which explains all their dependence in the sense of rendering them conditionally independent ($X \perp\!\!\!\perp Y \mid Z$). As a special case, Z may coincide with X or Y.*

The example of Nobel laureates X and chocolate consumption per capita Y illustrates this finding. While it is reasonable to assume that neither chocolate consumption makes Nobel laureates, nor does Nobel laureates make people eat more chocolate, a common cause Z may explain the observed correlation. Specifically, the prosperity of a country could drive the scientific success of its people, and enable more luxury goods (such as chocolate) to be consumed per capita. Therefore, the economic standing resembles a confounder Z, which causes both, more people becoming Nobel laureates, and more chocolate consumed.

Identifying the common cause requires background knowledge or additional assumptions, as it can not be passively observed. The observational distributions over X and Y can be explained in the case of X causing Y, Y causing X, and a common cause Z that causes both.

While correlation is useful in some scenarios, and causal inference in others, the field of medicine can profit from these insights. Treating a patient is a paradigmatic example of an intervention to influence a certain outcome. Causal inference can hence be used to develop more personalized medicine, to evaluate the efficacy of a treatment for sub-populations, and on a larger scale than an RCT allows. However, if we want to answer interventional questions, such as which treatment regiment and dosage are the most effective for this particular patient, more than just correlation is required: *a causal model.*

2.2 Causal Modeling Frameworks

There are multiple, co-existing approaches to causal modeling, of which two will be described in this section: the Causal Graphical Model (CGM), and the Potential Outcomes (PO) framework. First, the CGM is an illustrative framework that combines observed variable distributions with a directed graph, which offers an intuitive approach to causal inference. Second, the PO framework is popular in epidemiology and most appropriate for the medical application discussed in this thesis.

Causal Graphical Models (CGM) Directed graphical models are widely known as *Bayesian Networks* [25], which compactly represent joint probability distributions by graphically representing the dependencies between variables. But a causal interpretation of such a model requires a minor adaptation. When the edges in the directed, acyclic model determine the direction of the causal effect between two variables, we refer to them as CGMs [26].

Definition 4.1 (CGM). *A CGM $\mathcal{M} = (G, p)$ over n random variables $X_1, ..., X_n$ consists of: (i) a directed acyclic graph (DAG) G in which directed edges ($X_j \to X_i$) represent a direct causal effect of X_j on X_i; and (ii) a joint distribution $p(X_1, ..., X_n)$ which is Markovian w.r.t. G:*

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i | \mathbf{PA}_i), \tag{1}$$

where $\mathbf{PA}_i = \{X_j : (X_j \to X_i \in G\}$ denotes the set of parents, or direct causes, of X_i in G [9].

This is called the *causal (or disentangled) factorization* and is equivalent to the *Causal Markov condition*:

Definition 4.2 (Causal Markov condition). *A distribution p satisfies the causal Markov condition w.r.t. a DAG G if every variable is conditionally independent of its non-descendants in G given its parents in G.*

The causal edges of a CGM allow for determining the effect of interventions on an outcome variable. Generally, intervening on a variable means forcing it to take on a certain value. Thereby, the intervened upon variable is not caused by anything other than the intervention. Hence, it is graphically equivalent to deleting all the incoming edges to the variable. For instance, knocking out a gene means that it becomes independent of the regulatory mechanisms that caused its activation previously. The inactivity of the gene is now only caused by the intervention itself. Contrary, conditioning on the activity of the gene enables us to passively observe which conditions drive the activation of the gene.

Interventions were mathematically defined by Pearl's [27] *do-operator*. The notation $do(X = x)$ denotes the intervention that sets the RV X to the constant x. As this intervention removes all causes of it, the original graph G is modified s.t. the incoming edges into the intervened variable X are removed. This process was coined *graph surgery* [26]. The post-intervention graph G_M enables answering interventional queries using probabilistic inference. While CGMs are intuitive and offer a conceptually simple approach to interventional reasoning and inference, *counterfactual* reasoning can not be solved with them. For that purpose, Structural Causal Models (SCMs), or the Potential Outcomes (PO) framework can be utilized. This thesis will focus on the PO framework due to its popularity in epidemiology and the thesis' medical application.

Potential Outcomes (PO) The PO framework was first developed in randomized agricultural experiments [28], and later extended to observational studies [18]. Nowadays, it is used mostly in statistics and epidemiology. The popularity in epidemiology is unsurprising when considering the terminology: usually, the goal is to quantify the causal effect of a binary treatment variable T, where $T = 1$ and $T = 0$ indicate the treatment and control group respectively, on an outcome variable Y. This outcome variable is usually a measure of health.

Another interpretation of counterfactuals, which the PO framework tries to model, is to view them as missing data. The general notation of POs is $Y_i(t)$, which captures the outcome Y of individual i if they received treatment t. In the binary treatment case, one of the two POs is the observed factual, while the other is the unobserved counterfactual. The POs are considered fixed quantities for every individual i, and are therefore deterministic quantities. The randomness in the observed outcome Y arises from randomness in the treatment assignment:

$$Y_i = TY_i(1) + (1-T)Y_i(0). \tag{2}$$

The clinically meaningful *individual treatment effect* (ITE) can hypothetically be used to determine the treatment regiment of an individual i in the binary treatment condition is defined as:

$$\tau_i = Y_i(1) - Y_i(0). \tag{3}$$

Though, the "*fundamental problem of causal inference*" [19] states that one of these two POs remains the unobserved counterfactual:

$$Y_i^{CF} = (1-T)Y_i(1) + TY_i(0). \tag{4}$$

Therefore, τ_i is unidentifiable without further assumptions. In other words, it can not be directly computed from data [9].

The following two assumptions remained implicit in Equations 2 and 4:

Assumption 3.1 Stable unit treatment value (SUTVA). *The observation of one individual (i.e., unit) should be unaffected by the particular assignment of treatment to the other individuals [29].*

Assumption 3.2 Consistency. *If individual i receives treatment t, then the observed outcome is $Y_i = Y_i(t)$, i.e., the potential outcome for t [9].*

Assumption 3.1 means that (i) there is no interference between the individuals, and (ii) the treatment level is constant within the treatment and control groups, which leads to well-defined POs. For instance, if the treated group received a medication, it is important that all individuals received the same dosage (or treatment level) of this medication. Generally, this assumption enables us to view the individual units as independently sampled from the population.

Previously, POs were defined as deterministic quantities, although the study of complex subjects such as humans does not allow full characterization of the unit. Since this missing information induces uncertainty, POs are often defined as RVs. Generally, the confounders x_i of an individual i are observed and the expected POs are subject to reasoning by estimation of the expectation $\mathbb{E}[Y(1),\, Y(0) \mid x]$ from data.

Another assumption in the PO framework is that no confounders remain unobserved, which is equivalent to the Markov condition (Defn. 4.2) for CGMs. In the PO framework, this assumption of no hidden confounding between treatment and outcome is named *conditional ignorability*.

Assumption 2.3 Conditional Ignorability. *Given a treatment $T \in \{0,1\}$, POs Y(0), Y(1), and observed covariates X, which cause treatment and outcome, we have:*

$$Y(0) \perp\!\!\!\perp T \mid W \textit{ and } Y(1) \perp\!\!\!\perp T \mid W, \tag{5}$$

where $\perp\!\!\!\perp$ indicates the independence between Y and T given the adjustment set W. Conditional Ignorability is sometimes called Unconfoundedness because all relevant confounders to achieve conditional independence must be observed. If there are unobserved variables, which make T and Y dependent, the estimation of a treatment effect will be biased. Note that this is an untestable assumption since there may always be an unobserved third variable. As Assumption 2.3 suggests, the PO framework is targeted at investigating the (confounded) effect of some binary treatment on an outcome. Therefore, it is mostly used in causal reasoning, the main topic of this thesis.

2.3 Causal Reasoning

Causal reasoning aims at the quantification of causal relationships. This process requires two distinct steps: (i) the *identification* of the causal estimand with observable data; and (ii) the *estimation* of the causal effect using the data. The identification step aims at reducing a causal estimand to a statistical estimand. A causal estimand contains a *do-operator* and describes an intervention, while a statistical estimand is computable from observed conditional probabilities. In essence, identification entails selecting the variables to include in the estimation such that (s.t.) the result of an intervention can be computed based on conditional probabilities. Specifically, the graphical structure of the causal graph is investigated, and confounders that fulfill some graphical conditions are included in an adjustment set W with the aim of making T and Y conditionally independent. After selecting a sufficient adjustment set W, the statistical estimand is identifiable from the causal estimand and the causal effect can be estimated. Before the conditions for the adjustment set W are explained, let us define what we want to estimate. Generally, the causal effect consists of a difference in treatment effects and is estimated by contrasting two interventions.

Definition 3.1 Treatment effects. *The conditional average treatment effect (CATE) splits the sample based on an RV* $\mathbf{x}$*, and calculates a treatment effect for every subsample. Note that the RV* $\mathbf{x}$ *should be discrete as splitting a continuous variable will result in a uninformative abundance of subgroups. In case of an continuous RV* $\mathbf{x}$*, it is recommended to create meaningful ranges and converting the continuous variable to a discrete variable using those ranges. The CATE is defined as*

$$\tau(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{x}, do(T=1)] - \mathbb{E}[Y \mid \mathbf{x}, do(T=0)] = \mathbb{E}[Y(1) - Y(0) \mid \mathbf{x}]. \quad (6)$$

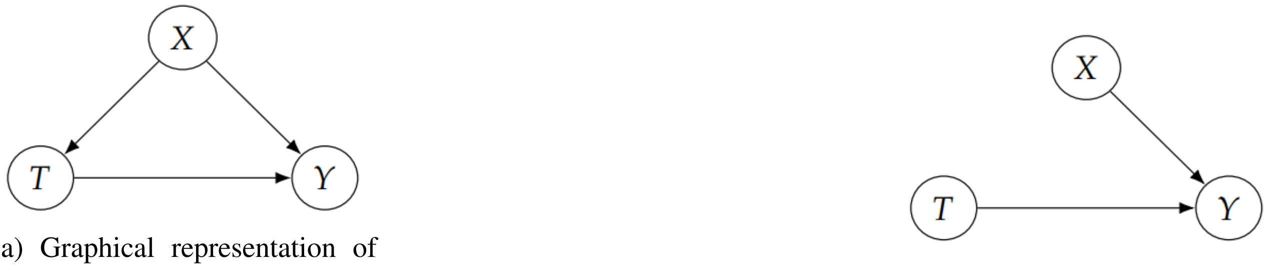
The average treatment effect (ATE) is, as the name suggests, defined as the population average of the CATE,

$$\tau = \mathbb{E}[Y \mid do(T=1)] - \mathbb{E}[Y \mid do(T=0)] = \mathbb{E}[Y(1) - Y(0)]. \quad (7)$$

There is a clear conceptual difference between the ITE (Equation 3), and the CATE (Equation 6). While the ITE refers to the treatment effect on the unit level, the CATE is the averaged treatment effect for a certain subpopulation, e.g. a Body-Mass-Index (BMI) above 30 in male individuals. While conceptually different, the CATE is used as an approximation of the ITE by calculating the CATE conditioned on the individual's features $\mathbf{x}_i$.

The *do-operator* in the Equations for the (C)ATE (Equations 6, 7) indicates that we aim at the estimation of interventional processes. Interventions are inherently different from simply conditioning on T. Conditioning entails restricting the available observational data to the comparison between sub-populations that did $(P(Y|T=1))$ or did not $(P(Y|T=0))$ receive the treatment. Contrary, interventions set the intervened upon treatment variable to a constant for the whole population [30]. This is possible if the treatment selection does not have any causes, which is graphically equivalent to deleting the incoming edges to T, and to conditional independence between T and Y. Achieving this state by selecting a sufficient adjustment set W to make T and Y conditionally independent is the goal of the identification step.

The notational differences are as follows. In conditioning, some variable T is observed to take a value t, so $P(Y=y|T=t)$ is the observed probability that $Y=y$ in the subset of the data where $T=t$. Contrary, the expression $P(Y=y|do(T=t))$ reflects the probability of $Y=y$ when T is set to the constant t. The latter reflects the population distribution of Y if the whole sample had the value of T fixed at t. The two notations can also be combined: $P(Y=y|do(T=t),X=x)$ describes the conditional probability of $Y=y$ in the subset of the data where $X=x$, in the distribution that results from the intervention $do(T=t)$ [8]. This results in a CATE estimate (Equation 6).



(a) Graphical representation of the state in observational data: The confounders influence the treatment selection.

(b) Graphical representation of the state in an RCT (Ignorability is satisfied).

Figure 1: Graphical representation of observational data and RCT data.

This intervention procedure affects the underlying probability function P. The original graph G (Figure 1a) with probability function P is manipulated to delete incoming edges to the treatment node T on which we want to intervene. The manipulated probability P_m is connected to the graph G_m (Figure 1b). While the manipulation changes P, two invariance relations hold. First, the marginal probability $P(X=x)$ remains unchanged because it does not have any parents, and it does not depend on T. Second, the conditional probability $P(Y=y|X=x,T=t)$ does not change because the deliberate manipulation of T does not change the process by which Y reacts to T [8]. The following holds:

$$P(X=x) = P_m(X=x)$$

and

$$P(Y|X=x,T=t) = P_m(Y|X=x,T=t)$$

Furthermore, since X is independent of T in the manipulated graph G_m, $P_m(X=x|T=t) = P_m(X=x) = P(X=x)$. Combining the above relations yields:

$$\begin{aligned} P(Y=y|do(T=t)) &= P_m(Y=y|T=t) && (8)\\ &= \sum_x P_m(Y=y|T=t,X=x)P_m(X=x|T=t) && (9)\\ &= \sum_x P_m(Y=y|T=t,X=x)P_m(X=x) && (10)\\ &= \sum_x \frac{P(T=t,Y=y,X=x)}{P(T=t|X=x)}. && (11) \end{aligned}$$

The right-hand side (RHS) of Equation 8 resembles the relationship we want to model by definition, i.e. the goal of causal inference is to evaluate the manipulated probability P_m of $Y=y$, given that $T=t$.

To achieve this, an adjustment set W is necessary, which makes T and Y conditionally independent. Specifically, this is an application of Bayes' rule. The process of identifying this adjustment set from all confounders is explained below. Equation 10 is called the adjustment formula and results from the fact that in the manipulated graph G_m, T, and X are independent. Furthermore, multiplying and dividing the summand in Equation 10 by the propensity score $P(T=t|X=x)$ yields Equation 11. Since Equations 10 and 11 contain only conditional probabilities, they can be computed from observational data [8]. Additionally, the conditional probability $P(T=t|X=x)$, called the propensity score, is sufficient to calculate the causal effect from the joint distribution $P(T=t, Y=y, X=x)$.

This relates to the general method of *g-computation*, which truncates the factorization defined in Equation 1 by the parent nodes of the intervened treatment variable. I.e., the product decomposition $P(X_1,...,X_n) = \prod_i P(X_i|PA_i)$ is truncated by excluding all nodes x_i that are in the set of intervention variables T. This leads to

$$P(X_1 = x_1, ..., X_n = x_n | do(T)) = \prod_{i \notin T} P(X_i|PA_i). \tag{12}$$

Identification The causal diagrams of real-world processes are never as simple as the toy graph in Figure 1a. Therefore, different criteria for the identification of a sufficient adjustment set W, s.t. T and Y are conditionally independent, were developed. This enables us to reduce a causal estimand, which is a formula that describes an intervention with the *do-operator* (LHS Equation 8), to a statistical estimand (Equation 11). The statistical estimand is a formula that contains only conditional probabilities, and can therefore be estimated from observational data. After the assumptions are met, and the statistical estimand is identified by using the sufficient adjustment set W, the remaining statistical association is causation.

One of these identification methods is called the *Backdoor criterion*. For a discussion of different methods see [8]. First, let us define the spurious association we strive to control for when identifying a causal estimand.

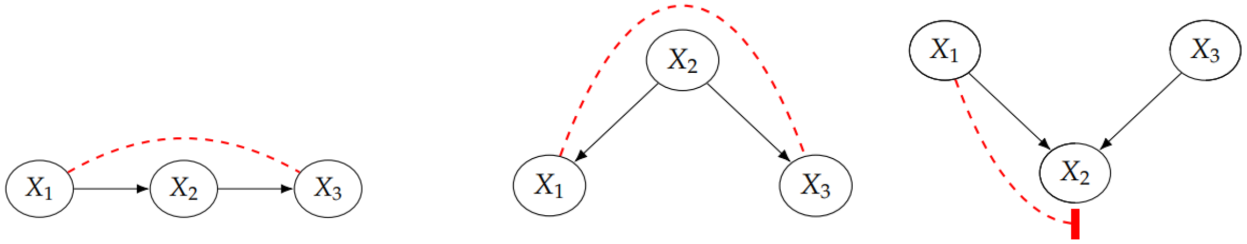
Spurious association flows through chains and forks that are not conditioned on, and through colliders, if they are conditioned on [31]. Figure 2 shows three DAGs, the nodes of which are RVs. In Figure 2a, the variables X_1 and X_3 are statistically dependent through X_2, and conditioning on the chain node X_2 would make them statistically independent. Similarly, in Figure 2b, X_1 and X_3 are statistically dependent and can be made independent by conditioning on the fork node X_2. In contrast, X_1, and X_3 are statistically independent in Figure 2c because node X_2 is a collider. In other words, the statistical dependence between X_1 and X_3 is blocked by node X_2. They are made statistically dependent by conditioning on the collider X_2. This can be summarized in 3 rules:

1. **Conditional Independence in Chains** Two variables, X_1 and X_3, are conditionally independent given X_2, if there is only one unidirectional path between X_1 and X_3 and X_2 is any set of variables that intercepts that path. ([8], p. 39).

2. **Conditional Independence in Forks** If a variable X_2 is a common cause of variables X_1 and X_3, and there is only one path between X_1 and X_3, then X_1 and X_3 are independent conditional on X_2. ([8], p. 40).

3. **Conditional Independence in Colliders** If a variable X_2 is the collision node between two variables X_1 and X_3, and there is only one path between X_1 and X_3, then X_1 and X_3 are uncon-

ditionally independent but are dependent conditional on X_2 and any descendants of X_3. ([8], p. 44).



(a) Association flows from X_1 to X_3 through the chain created by X_2

(b) Association flows from X_1 to X_3 through the fork created by X_2.

(c) A collider blocks the flow of association.

Figure 2: An overview of the flow of association in DAGs.

These conditional independence relations are combined in the *Backdoor criterion*, a selection of graphical conditions that, when fulfilled, enable the identification of the computable, statistical estimand from the causal estimand. This conditioning set W has to consist of RVs of the Causal Graph, such that the conditions in the Backdoor criterion (Definition 3.1) are satisfied. The set W consists of the common causes for both T and Y. Graphically, conditioning on W is equivalent to deleting the incoming edges to T (see Figure 1b). Additionally, the Positivity assumption (Assumption 3.1) must be satisfied.

Definition 3.2 Backdoor Criterion. In a DAG G with an ordered pair of variables (T, Y), the backdoor criterion is satisfied by a set of variables W that adheres to the following [8]:

1. W blocks all paths between the ordered pair (T, Y) that contains an arrow into T, and is not the direct arrow from T into Y. This blocking is achieved by conditioning a variable along the path.
2. W does not contain any descendants of T.

Assumption 3.1 Positivity. For all subgroups of the data with a certain level of confounding (i.e. if P(W=w) > 0), the probability of being selected for treatment can not be 0 or 1 (i.e. $0 < P(T = 1|W = w) < 1$).

Now let us illustrate the necessity of causal assumptions using a popular statistical phenomenon.

Simpson's Paradox and Covid-19 Simpson's paradox refers to the observation that aggregating data on subpopulations can lead to opposite trends (and thus opposite conclusions) when subpopulations are considered separately [32]. We observed an excellent example of this during the Covid-19 pandemic when we combined case fatality rates (CFRs), i.e., the proportion of confirmed Covid-19 cases that are fatal, across different countries and age groups, as shown in Figure 3 [33]: For all age groups, CFRs are *lower* in Italy than in China, but the overall CFR in Italy is *higher*.

How can this pattern be explained? The demographics of the cases (see Figure 3, right) are quite different in the two countries, i.e., there is a statistical correlation between country and age. Italy, in particular, had a much higher proportion of cases in older patients, who generally have a higher risk of dying from Covid-19 (see Figure 3, left). While this explains the phenomenon statistically, it may seem puzzling, as it defies causal intuition. Humans seem to naturally extrapolate conditional probabilities to be interpreted as causal effects, which can lead to contradictory conclusions, and

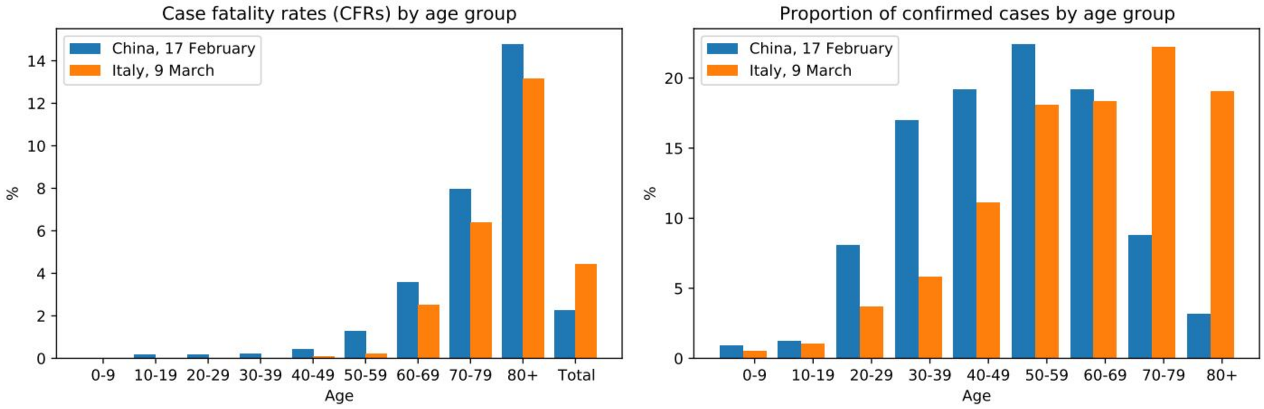


Figure 3: Left: Case fatality rates (CFRs) of Covid-19 in China and Italy. It includes cases reported until early 2020 (see legend). For every group, CFRs in Italy are lower compared to China but the aggregated CFR is higher in Italy. This is an example of *Simpson's paradox*. Right: The case demographics for China and Italy. While in China most cases were recorded for 40-49 year-olds, in Italy, the two oldest age groups (70-80+) were the most prevalent. This figure is from [33].

raises the question: *How is it possible that the disease in Italy is less deadly for the young, less deadly for the elderly, but more deadly for the people as a whole?* For this reason, the inversion of the probabilities in Figure 3 is called a ”paradox” [34, 35].

If one considers the country as a treatment whose causal effect on mortality is of interest, then a causal diagram is necessary to decide how to treat covariates, such as age, that are statistically associated with the treatment, e.g., whether to stratify (i.e., adjust) for age or not. This also explains why RCTs [36] are the gold standard for causal inference: randomization ensures that the confounding variables can not have an outcome on the treatment variable, i.e., someone's age does not determine whether or not they (do not) receive the treatment. This ensures that there is no potential bias. However, RCTs are costly and sometimes unethical, so causal inferences are usually based on observational data only.

3 Methods

The following section serves multiple purposes. First, the estimators used to compute the causal effects are introduced. Second, an extension to one of these estimators, the X-learner++, is described. Third, the semi-synthetic dataset, which enables us to verify the estimator performance, is explained. Then, the medical question this thesis aims to answer is stated. Last, we describe the medical dataset under investigation.

3.1 Estimators

This section discusses different estimators for estimating causal effects from observational data. The chosen estimators are Inverse Probability Weighting, and Conditional Outcome Estimation, which is combined in a doubly robust estimator. Furthermore, the X-Learner and debiased Machine Learning (ML) are discussed.

3.1.1 Inverse Probability Weighting (IPW)

The concept of truncation (Equation 12) becomes apparent when comparing the formula for the pre-intervention distribution $P(Y=y,T=t,X=x)$ (Equation 13) of Graph G in Figure 1a to the adjustment formula. Equation 13 differs from Equation 10 only with respect to the term $P(T|X=x)$.

$$P(Y=y,T=t,X=x)=\sum_{x}P(Y|T=t,X=x)P(T|X=x)P(X=x) \tag{13}$$

After deleting the incoming edge to T in the manipulated Graph G_m in Figure 1b, T does not have any parents. In other words, in this manipulated Graph, nothing causes T. This simple relationship between the pre-intervention probability distribution P, and the post-intervention probability distribution P_m enables us to calculate P_m given a certain intervention by multiplying Equation 13 by the inverse $\frac{1}{P(T|X=x)}$ of the propensity score (Equation 11).

The relationship between the pre- and post-intervention probability distribution, P and P_m respectively, has multiple advantages. While the adjustment formula works reliably, increasing the dimensionality of the adjustment set W may bear problems. When W consists of multiple variables, that all can take many values, the summation of all values of W encounters computational and estimation difficulties. For instance, the data in specific strata $W=w$ may be too small to allow for reliable estimation. Given a reliable estimate of the function $P(T=t|W=w)$, and given that W satisfies the Backdoor criterion, the available data can be re-weighted to act as a sample drawn from the post-intervention distribution P_m. This artificial data can then be used to evaluate $P(Y|do(t))$ by counting frequencies for $Y=y$ in the sample where $T=t$.

This method is best understood using an example, which follows [8] and is taken from [37]. In this example, Y is a binary RV that indicates whether a patient recovered or not, the binary RV T indicates if a drug was administered, and the RV W indicates the gender of the patient. We assume that W satisfies ignorability and that the propensity score $P(T|W)$ satisfies positivity.

T	Y	W	% of population
Yes	Yes	Male	0.116
Yes	Yes	Female	0.274
Yes	No	Male	0.01
Yes	No	Female	0.101
No	Yes	Male	0.334
No	Yes	Female	0.079
No	No	Male	0.051
No	No	Female	0.036

Table 1: An example of a study that investigates the effect of a treatment on recovery based on gender. T indicates if the drug was administered, Y if the patient recovered, and W indicates the gender of the patient.

To estimate $P(Y|do(T = yes))$ from the data, the propensity score $P(T|W = w)$ for each value of W is calculated. Based on Table 1, this results in the following calculation:

$$P(T = yes|W = Male) = \frac{0.116+0.01}{0.116+0.01+0.334+0.051} = 0.233$$

$$P(T = yes|W = Female) = \frac{0.274+0.101}{0.274+0.101+0.079+0.036} = 0.765$$

The probabilities calculated above are now used to re-weight the corresponding gender rows in the strata that did receive the treatment ($T = Yes$) by the inverse of these probabilities. Specifically, rows 1 and 3 are weighted by $\frac{1}{0.233}$, while rows 2 and 4 are weighted by $\frac{1}{0.765}$. This probability distribution in Table 2 reflects the post-intervention distribution P_m. Based on the data, the post-intervention probability of recovery can now be computed. If everyone in the population was given the drug ($T = Yes$), the probability of recovery is:

$$P(Y = yes|do(T = yes)) = 0.476+0.357 = 0.833$$

T	Y	W	% of population
Yes	Yes	Male	0.476
Yes	Yes	Female	0.357
Yes	No	Male	0.041
Yes	No	Female	0.132

Table 2: The post-intervention distribution of the intervention do(T=Yes) in the population in Table 13. The distribution was determined using the IPW method.

IPW also allows us to calculate the ATE. This is done in two steps. First, the observations in the sample that did get the treatment $T = 1$ are re-weighted by the inverse of the propensity score $P(T|W)$.

Similarly, the observations that did not get the treatment $T = 0$ are multiplied by the inverse of $1 - P(T|W)$. Second, the empirical mean is calculated and the two results are subtracted from one another. This results in Equation 14 [17].

$$A\hat{T}E_{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{X_iY_i}{\hat{P}(T|W)} - \frac{(1-X_i)Y_i}{(1-\hat{P}(T|W))} \tag{14}$$

This estimator suffers from poor sample properties when the propensity score approaches 0 or 1 for observations in the sample. If an observation i in the sample received the treatment (T=1) and has a propensity score close to 0, the contribution of this data point will be more extreme, to the point where the contribution goes beyond the possible range for the ATE. This can be fixed by renormalizing the propensity scores such that they sum up to 1 [38, 39]. This results in Equation 15.

$$A\hat{T}E_{IPW*} = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{\hat{P}(T|W)}\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \frac{X_iY_i}{\hat{P}(T|W)} - \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1-X_i}{1-\hat{P}(T|W)}\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \frac{(1-X_i)Y_i}{(1-\hat{P}(T|W))} \tag{15}$$

3.1.2 Conditional Outcome Estimation (COM)

Another approach to estimating the ATE is to fit a machine learning model to estimate the conditional expectation $\mathbb{E}[Y|do(T=t),W]$ as a function of W, by taking the empirical mean $\hat{\mu}$ over all data points n. This assumes that W is a sufficient adjustment set that satisfies the backdoor criterion. This results in the Equation

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}(do(T=1),w_i) - \hat{\mu}(do(T=0),w_i)). \tag{16}$$

Specifically, one single model is fit to regress the outcome Y on the treatment T and the set of confounders W on the original dataset. Thereafter, the data is intervened upon. In the binary treatment case $T \in [0,1]$, this creates two data sets that differ only in their value for T. The model trained on the original data set is used to predict the two intervened data frames to estimate $\mathbb{E}[Y|do(T=1)]$ and $\mathbb{E}[Y|do(Y=0)]$. Since the only difference between the dataset is the value for T, the model may ignore this subtle difference in the case of a high-dimensional adjustment set W. This would bias the treatment effect towards zero since the results of both predictions is equal if the model ignores the difference in the treatment variable T in the two data sets. This may be alleviated by fitting two separate models, $\hat{\mu}_0$ and $\hat{\mu}_1$. Each model is only trained with the part of the data where $T = 0$ and $T = 1$, respectively. This is called grouped COM estimation (GCOM). This approach suffers from data inefficiency since only part of the data is used to train the models. Also, the more dimensions W has, the more likely it gets that there is a lack of overlap between the part of the data that did get the treatment, compared to the part that did not get the treatment. This may induce uncertainty in the estimator and finite sample bias [40].

3.1.3 AIPW

The IPW and (G)COM estimators may both be misspecified. This means that the model has biased coefficients and error terms. To alleviate this shortcoming, both estimators can be combined. As a result, only one of the two estimators has to be specified correctly to make the result robust. For an explanation of why this is the case, see [12], p.178. This yields the doubly robust AIPW estimator in Equation 17 [12].

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i(Y_i - \hat{\mu}_1(W_i))}{\hat{P}(T|W)} + \hat{\mu}_1(W_i)\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-T_i)(Y_i - \hat{\mu}_0(W_i))}{1-\hat{P}(T|W)} + \hat{\mu}_0(W_i)\right) \tag{17}$$

Where Y is a continuous outcome variable, T is a binary treatment variable, and $\hat{\mu}_1$ is the regression model trained on the part of the data where $T = 1$, $\hat{\mu}_0$ for the part where $T = 0$, and $\hat{P}(T|W)$ is the estimation of the propensity score. If sample i was treated, the residuals between Y_i and $\hat{\mu}_1(W_i)$ are re-weighted by the estimated propensity score. Afterwards, the estimated value $\hat{\mu}_1(W_i)$ is added to the re-weighted residuals, and the empirical mean over all i is computed.

Assuming ignorability and positivity, the bias of this estimator is small if either the GCOM or the propensity score model is specified correctly. Specifically, for the first part of the equation, the bias in large samples is

$$\mathbb{E}[P(T|W)\left(\frac{1}{P(T|W)} - \frac{1}{P^*(T|W)}\right)(\mu_1(W) - \mu_1^*(W)]. \tag{18}$$

Where $P^*(T|W)$ and μ_1^* represent the probability limits of the models. If the models are correctly specified, $P^*(T|W)$=$P(T|W)$ and μ_1^*=μ_1. If one of the models is correctly specified, then either $\frac{1}{P(T|W)} - \frac{1}{P^*(T|W)} = 0$ or $\mu_1(W) - \mu_1^*(W) = 0$, which results in the whole Equation 18 to be 0 due to multiplication with 0. Hence, if either one of the models is correctly specified, the asymptotic bias is zero [12].

3.1.4 X-Learner

The X-Learner [16] improves upon the data inefficiency of GCOM estimation by adding two more steps. In the first step of the algorithm, two models $\hat{\mu}_0$ and $\hat{\mu}_1$ are trained to estimate the mean outcomes in the portion of the data where $T = 0$ and $T = 1$ (Equations 19 and 20, respectively). These are called the base learners of the first stage [16] and are equivalent to the GCOM estimators.

$$\hat{\mu}_0 = \hat{\mathbb{E}}[Y|T = 0, W] \tag{19}$$

$$\hat{\mu}_1 = \hat{\mathbb{E}}[Y|T = 1, W] \tag{20}$$

These models are then used to estimate the ITEs (Equation 3). To do this, the counterfactual values are imputed using the relevant base learner of the first stage. That is, for every individual in the portion of the sample where $T = 1$, the ITE is estimated using Equation 21. This is done by taking the observed, factual value for Y, and predicting what the counterfactual outcome for Y would have been if the individual had not received the treatment. The counterfactual is imputed using the model $\hat{\mu}_0$, which was trained on the portion of the data where $T = 0$. For every individual where $T = 0$, the ITE is estimated using Equation 22. Here, the counterfactual is predicted by using the model $\hat{\mu}_1$, which was trained on the data where $T = 1$ to predict the counterfactual in the portion of the data where $T = 0$.

$$\tau_b = Y_b - \hat{\mu}_0(W_b) \tag{21}$$

$$\tau_j = \hat{\mu}_1(W_j) - Y_j \tag{22}$$

The b in Equation 21 indicates the portion of the data where $T = 1$, while j in Equation 22 stands for the portion of the data where $T = 0$. The next step is to train two more models coined as the base learners of the second stage. First, the ITEs from Equation 21 are the response variable for a model $\hat{\tau}_1(W)$, based on the treatment group data (where $T = 1$). Similarly, a model $\hat{\tau}_1(W)$ uses the results from Equation 22 as the response variable and the confounders from the control group data (where $T = 0$) as predictors. All models can be trained using any supervised learning algorithm that can cope with non-linearity and high dimensionality.

Finally, the two second-stage base learners are combined to give a single CATE estimate. This is done by weighting both base learners of the second stage by a function $g(x) \in [0,1]$. A suitable weighting function is the propensity score [16], which yields

$$\hat{\tau} = \hat{P}(T|W)\hat{\tau}_0(w) + (1 - \hat{P}(T|W))\hat{\tau}_1(w). \tag{23}$$

3.1.5 X-Learner++

This thesis investigates an extension to the X-learner, which corresponds to an error correction and was invented in this thesis. Specifically, the error in estimating the counterfactuals from the base learners is quantified. After the counterfactuals are imputed using the models trained in Equations 19 and 20, two models $\hat{e}_0$ and $\hat{e}_1$ can be trained to predict the imputed counterfactuals $\hat{\mu}_0$ and $\hat{\mu}_1$ from the confounders W, where $T = 1$ and $T = 0$, respectively.

$$\hat{e}_0 = \hat{\mathbb{E}}[\hat{\mu}_0(W_i)|T = 1, W] \tag{24}$$

$$\hat{e}_1 = \hat{\mathbb{E}}[\hat{\mu}_1(W_i)|T = 0, W] \tag{25}$$

For example, the model $\hat{e}_0$ is trained on the part of the data where $T = 1$ to predict the outcome of the model $\hat{\mu}_0(W_i)$. I.e., the model $\hat{e}_0$ learns to predict the counterfactual $Y_{cf}(0)$ for the confounders where $T = 1$. Thereafter, the model is used to predict the factual $Y(0)$ from the confounders W where $T = 0$. This prediction now yields values that can be subtracted from the observed factual to yield an indication of the error in predicting the counterfactuals. This error is halved and added to each counterfactual prediction to correct the error. The halving is because the error should occur twice, once in the counterfactual prediction from the base learners, and once in the factual prediction of this error correction.

$$E_{0i} = \frac{Y(0)_i - \hat{e}_0(W_i|T = 0))}{2} \tag{26}$$

$$E_{1i} = \frac{(Y(1)_i - \hat{e}_0(W_i|T = 1))}{2} \tag{27}$$

Now, the Equation for the estimation of the ITEs turns from Equation 21 and 22 into

$$\tau_i = Y_i - (\hat{\mu}_0(W_i) + E_{0i}), \tag{28}$$

and

$$\tau_i = (\hat{\mu}_1(W_i) + E_{1i}) - Y_i. \tag{29}$$

The rest of the estimator remains the same. This error correction makes full use of the information available to quantify the mistakes made in the counterfactual prediction.

3.1.6 Debiased Machine Learning

In causal inference, we strive to estimate the treatment effect from a high-dimensional set of variables consisting of confounders and the treatment variable. Furthermore, the reduction of bias in this estimation is desirable. To this end, the Frisch-waugh-Lovell (FWL) theorem[1] [42] is combined with flexible ML methods to yield the debiased ML estimator. The FWL theorem reduces bias in the estimation while the flexible ML methods are well suited to estimate high-dimensional data. To reduce the risk of overfitting the ML models, cross-fitting is used. The combination of those two approaches makes the estimator $N^{-\frac{1}{2}}$ consistent, where N is the sample size.

Importantly, the purpose of cross-fitting is different from cross-validation, even though the methods are similar. In k-fold-cross-fitting, the dataset is split into *k* parts. Then, for each split, a model is trained on the *k-1* subsets of the data. This model is subsequently used to predict the out-of-fold subset. This means that the ML models never predict any data they have encountered before, which reduces the chance of overfitting. Note that in cross-validation, the out-of-fold predictions are used to attain an unbiased estimate of how well the model performs. However, in cross-fitting, we solely strive to get an unbiased prediction with a low likelihood of overfitting.

Debiased ML was developed to handle semi-parametric inference of a low-dimensional parameter θ_0 from high-dimensional nuisance parameters η_0 [15]. The method combines the Frisch-Waugh-Lovell (FWL) theorem[2] [42] with flexible ML methods, which are well suited to estimate high-dimensional data. Generally, the FWL theorem and cross-fitting are used to reduce the bias induced by regularization and overfitting of modern ML methods, which may hinder the estimator to be $N^{-\frac{1}{2}}$ consistent, where N is the sample size.

Frisch-Waugh-Lovell Theorem Frisch, Waugh, and Lovell are credited[3] with the discovery of an interesting property of partial linear regression. This property is best demonstrated with an example. Assume a linear regression model with two feature matrices, X_1 and X_2, then:

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2,$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are row vectors that must be estimated to predict Y from the features X_1 and X_2. Instead of estimating β_1 directly, the following steps lead to the same results:

1. Regress Y on the second set of features X_2:
$$\hat{Y}^* = \hat{\gamma}_1 X_2.$$

2. Regress X_1, the first set of features on the second set of features X_2:
$$\hat{X}_1 = \hat{\gamma}_2 X_2.$$

[1] see [41] for a short proof of the theorem.

[2] see [41] for a short proof of the theorem.

[3] Even though G. U. Luwe published the theorem 26 years prior [43] (section 9, page 184).

3. Calculate the residuals:

$$\tilde{X}_1 = X_1 - \hat{X}_1,$$

and

$$\tilde{Y} = Y - \hat{Y}^*.$$

4. To yield $\hat{\beta}_1$, regress the outcome residuals $\tilde{Y}$ on the features residuals $\tilde{X}_1$:

$$\tilde{Y} = \hat{\beta}_1 \tilde{X}_1.$$

This generic case can be specified to suit the problem statement in causal inference by reducing the set of features X_1 to a single feature, the treatment variable T. In other words, to derive the treatment effect, regression on residuals can be performed to yield the low-dimensional treatment parameter τ:

$$(Y - (Y \sim W)) \sim (T - (T \sim W)),$$

where $\sim$ indicates a regression model. Or, equivalently

$$Y - \mathbb{E}[Y|W] = \tau(T - \mathbb{E}[T|W]) + \varepsilon,$$

where Y is the RV corresponding to the outcome, T is the treatment variable, and W is the set of confounders that satisfies the *Backdoor criterion* (Definition 3.9). In essence, the AWL theorem allows us to split the estimation of confounders to the treatment from the estimation of the causal parameter τ. When the models are fitted with enough data, the result should be an unbiased estimate of the treatment effect.

Debiasing with ML This Section follows [42]. Modern ML models are efficient function approximators. Their flexibility is well-suited to estimate the expectations mentioned in the equation above, yielding

$$Y - \hat{M}_y(W) = \tau(T - \hat{M}_t(W)) + \varepsilon,$$

where $\hat{M}_y$ and $\hat{M}_t$ are ML models that estimate $\mathbb{E}[Y|W]$ and $\mathbb{E}[T|W]$, respectively. After calculating the residuals, the causal parameter τ can be estimated using a linear model, such as ordinary least squares (OLS):

$$\tilde{Y} = \alpha + \tau \tilde{T}.$$

While the flexibility of ML models helps deal with the function approximation of non-linear, high-dimensional nuisance parameters, it bears the risk of overfitting. In essence, an overfitted model leads to a reduction in the variance of the residuals in either the outcome residuals $\tilde{Y}$ or the treatment residuals $\tilde{T}$. Consequently, if M_y or M_t are overfitted, the model captures the treatment effect or makes it difficult to compare treatment levels respectively (see [15] for a rigorous explanation of the problems of overfitting in this context). This issue can be overcome by using cross-fitting as described above.

Following this procedure guarantees an asymptotically normal estimator, which allows the computation of valid confidence intervals. Furthermore, the estimator is approximately unbiased, and $N^{-\frac{1}{2}}$ consistent. For a more in-depth explanation with empirical examples, see [15].

3.1.7 Ensemble

Another contribution of this thesis is the investigation of an ensemble of the four estimators. The ensemble ATE is calculated by taking the arithmetic mean μ over all four results.

3.2 Estimation

The goal of this thesis is to estimate the ATE for two datasets. First, a semi-synthetic dataset based on the Infant Health Development Program (IHDP). The IHDP is a randomized, clinical trial that evaluates the efficacy of an early treatment to reduce the developmental and health problems of low birth weight, and premature infants [20]. This covariate data was used to synthesize the outcomes such that the evaluation of causal inference estimators is possible [44]. Second, a real-world dataset that evaluates the efficacy of chemoprophylaxis for VTE after an ankle fracture is investigated. The first dataset is meant to show the efficacy of the estimators at hand, while the second dataset resembles one of the first applications of the causal inference methodology in the field of orthopaedics. For both datasets, the best-performing models to estimate the outcome and the propensity score were identified using a grid search approach as outlined in Section 4.2.

3.2.1 Semi-synthetic Dataset

The IHDP dataset resembles a semi-synthetic dataset that has real-world complexities, while the outcomes are simulated. Therefore, access to the ground truth for the ATE is available. Since the original dataset stems from the IHDP RCT from 1985, ignorability is satisfied and a causal diagram is not necessary [44]. The IHDP experiment investigated the effect of high-quality child care and home visits from a trained provider for infants that suffer from low birth weight or are born prematurely. The measured outcome was cognitive test scores, which showed a large effect of the intervention in the real dataset [45].

In the semi-synthetic dataset, there are 19 binary, and 6 continuous covariates in total [44]. For instance, there are child measurement variables such as birth weight, head circumference, weeks born preterm, birth order, firstborn, neonatal health index, sex, and twin status. Furthermore, pregnancy behaviors are included such as whether cigarettes were smoked, alcohol was drunk, or drugs were consumed. Last, descriptive variables of the mother at the time of birth were recorded. These include age, maritial status, educational attainment, whether she worked during pregnancy, and whether she received prenatal care.

To simulate an observational study from the experimental data, a non-random portion of the data was discarded [44]. Specifically, the children with non-white mothers were excluded from the treatment group. As a result, the treatment and control groups are not balanced. Thus, simple outcome comparisons would lead to biased estimates, whereas the estimators described above should be able to estimate the correct ATE. After this exclusion, there are 139 children in the case group and 608 children in the control group. Table 3 shows an overview of the class distribution in both datasets under investigation.

For this dataset, the best model to predict the outcome was the `Gradient Boosting Regressor` [46] with a learning rate of 0.01, a maximum depth of 2, and 1000 estimators. For the propensity score, the best-performing model was the `AdaBoost Classifier` [47] with a learning rate of 0.4 and 500 estimators.

	Dataset			
Groups	IHDP		VTE	
Case	139	18.6%	239	20.3%
Control	608	81.4%	936	79.7%
Total	747	-	1175	-

Table 3: The distribution of treatment vs. control group for the ATE estimation in both, the IHDP and the VTE dataset.

3.2.2 Medical Problem Statement

VTE is a leading cause of death after major orthopaedic surgery. It is a combination of two disorders: deep vein thrombosis (DVT) and pulmonary embolism (PE), both of which can be lethal. The Surgeon General to the United States of America (USA) declared VTE a public health concern since up to 600,000 people die each year in the USA [48].

The condition is difficult to detect in its early stages, and surgeons disagree on when to administer chemoprophylaxis [49], i.e., the administration of anticoagulants, which makes the formation of blood clots less likely but increases the chance of severe bleeding [50]. The disagreement is especially apparent in isolated foot and ankle fractures. Specifically, patients not at risk of VTE are suggested not to receive chemoprophylaxis due to the possible risk of bleeding adverse events (BAEs) [51, 52].

This thesis investigates the general effect of chemoprophylaxis on VTE incidence, and the conditional average treatment effect, conditioned on prior use of Statins. Statins are the primary treatment for hyperlipidemia, an increased level of fats in the blood. Many patients with cardiovascular disease (cvd) consume Statins already. Furthermore, Statins are suspected to lower the risk of VTE, as shown in two RCTs [53, 54]. However, the effect of Statins is not immediate and the combined effect of chemoprophylaxis and Statin consumption remains unexplored. Therefore, the effect of Statins on the efficacy of chemoprophylaxis is an open field of research.

3.2.3 Medical Dataset

The data was gathered using Mass General Orthopaedic Registry and the research patient data registry tool (RPDR). A total of 16,421 patients with ankle fractures were recruited retrospectively. All these patients were visited in one of the three hospitals in the Mass General Brigham network including Massachusetts General Hospital, Brigham and Women Hospital, and Newton Wellesley Hospital, between January 2004 and June 2021. To hasten the process of screening the patients' data an automated string search method was used to find patients who were suspect VTE or VTE was mentioned in their records within 180 days after the ankle fracture. This resulted in 1,175 patients. Out of 1,175 patients, 239 had confirmed VTE 180 days after the incidence of ankle fracture. The control group consists of 936 patients. Table 3 shows the class distribution. The inclusion criteria were:

1. Presence of an ankle fracture diagnosed by a physician and confirmed radiologically via X-ray or CT scan
2. Age of 18 years or older
3. Symptomatic VTE confirmed by a clinician and through radiologic (Duplex ultrasound, CT angiography, and/or angiography) and laboratory (D-Dimer) assessments

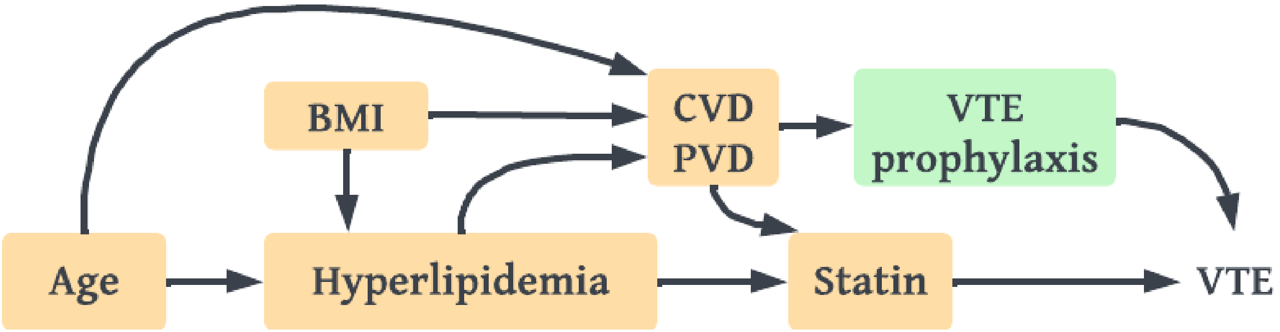


Figure 4: The causal diagram of the interaction between Statins, VTE chemoprophylaxis, and VTE. The treatment is colored green, while the included variables are colored yellow.

The exclusion criterion is:

1. Patients who did not have a confirmed and symptomatic VTE or VTE occurred after 180 days post-fracture.

All data and notes of the patients in every hospital encounter were manually screened via a structured chart review and clinical documentation in order not to miss any findings or symptoms indicating VTE or any complication that could affect the outcomes of our patients. Furthermore, the patients in each group were split into two subgroups, one of which already consumed Statins, while the other did not. An overview of the class distribution for the CATE estimation is presented in Table 4. Missing data were imputed using the Multivariate Imputation by Chained Equations (MICE) technique [55].

For this dataset, the best model to predict the outcome was the `AdaBoost Classifier` [47] with a learning rate of 0.1 and 500 estimators. For the propensity score, the best-performing model was the `Random Forest Classifier` with a maximum depth of 50 and 5 estimators.

	Dataset			
Groups	Statins		no Statins	
Case	50	11.2%	542	74.1%
Control	395	88.8%	189	25.9%
Total	445	-	731	-

Table 4: The distribution of treatment vs. control group for the CATE estimation in the VTE dataset.

To identify the variables necessary to satisfy the Backdoor criterion (Definition 3.1), physicians at the SORG and FARIL collaborative research lab at Harvard Medical School conducted an internal systematic review. This review aimed at identifying the causal structure of the interaction between Statins, VTE chemoprophylaxis, and VTE. The result of this review is displayed in Figure 4. While Age and BMI would already be a sufficient adjustment set, including all variables increased the predictive power of the model. Therefore, all variables were included in the analysis.

3.2.4 Cross-Fitting and Boostrapping

Each estimator under investigation in this thesis is asymptotically normal. Therefore, valid Confidence intervals (CIs) can be constructed using bootstrapping. Hence, 5,000 samples were created from the original dataset by sampling with replacement. Then, the (C)ATE is calculated on every sample. Furthermore, k-fold cross-fitting was implemented and used for the estimation of the data. This is recommended as best practice to reduce the bias in the estimation of the causal effect [15, 56].

4 Experimental Setup

This Section describes the specific steps taken in the experiment process.

4.1 Tools and Technologies

Machine learning models and k-fold data splitters were imported from the scikit-learn library [57]. Also, the basic standardizer *scale* was used to standardize the continuous variables after sample splitting. Furthermore, the set of estimators was run on a Windows 11 Desktop computer with 32GB RAM, an *AMD Ryzen 9 5900 12-Core Processor*, and a *NVIDIA GeForce RTX 3080*.

4.2 Experimental Configurations and Hyperparameter Optimization

The model selection and hyperparameter optimization for the ML models used in the estimation step were combined in one large grid search. The models and hyperparameters under investigation are displayed in Appendix Section B. All experiments were performed using the random seed 123 to allow for reproducibility.

4.3 Performance Criteria

During the grid search, multiple scoring systems were used, depending on the predicted outcome. For binary outcomes, the weighted F1-score was used to evaluate model performance. For instance, the best propensity score model was selected based on the performance of classifying the binary treatment variable. In the estimation step, however, `predict_proba` is used to attain a continuous prediction from a classifier. For the continuous outcomes, the negative mean squared error indicated model performance. Note that these metrics were used for model selection and that the hyperparameters of each model were optimized using k-fold cross-validation before comparing the performance. Ultimately, the target metric under investigation in this thesis is the (C)ATE.

For the semi-synthetic IHDP dataset, the synthesized outcomes enable a direct evaluation of the estimator's performance. Hence, the error between the estimated ATE and the ground-truth ATE is reported (Equation 30), where ATE_{Est} is the estimated ATE and ATE_{GT} is the ground truth ATE.

$$Err_{ATE} = \overline{ATE}_{Est} - \overline{ATE}_{GT} \tag{30}$$

There is no ground truth available for the medical dataset. Therefore, the performance of the estimations based on the VTE dataset is evaluated based on a subsample of a meta-analysis that investigated the effect of VTE chemoprophylaxis on the VTE incidence based on 22 studies [51]. This subsample of $1,666$ patients had radiologically confirmed VTE, like in the dataset studied in this thesis. The rest of the meta-analysis contained patients that were only clinically confirmed and is hence not as comparable.

5 Results

5.1 Model Selection

The model selection scores for the outcome model in the IHDP dataset are presented in Table 5, while Table 6 contains the results for the IHDP propensity score model.

Model	negative MSE
Gradient Boosting Regressor	-0.37
AdaBoost Regressor	-0.41
Random Forest Regressor	-0.87
SVM	-1.53
MLP	-1.7

Table 5: This Table shows the performance of the five best models when predicting the outcome of the IHDP dataset.

Model	weighted F1-score
AdaBoost Classifier	0.87
Gradient Boosting Classifier	0.84
Random Forest Classifier	0.79
Logistic Regression	0.79
Bagging	0.72

Table 6: This Table shows the performance of the five best models when predicting the treatment variable of the IHDP dataset.

The model selection scores for the outcome model in the IHDP dataset are presented in Table 5, while Table 6 contains the results for the IHDP propensity score model.

The results for the model selection in the VTE dataset are presented in Table 7 for the outcome model, and in Table 8 for the treatment prediction.

Model	weighted F1-score
AdaBoost Classifier	0.81
Random Forest Classifier	0.79
Bagging	0.73
Logistic Regression	0.73
Gradient Boosting Classifier	0.72

Table 7: This Table shows the performance of the five best models when predicting the outcome of the VTE dataset.

Model	weighted F1-score
Random Forest Classifier	0.77
AdaBoost Classifier	0.76
Stacking	0.73
Bagging	0.71
Gradient Boosting Classifier	0.7

Table 8: This Table shows the performance of the five best models when predicting the treatment variable of the VTE dataset.

5.2 IHDP Dataset

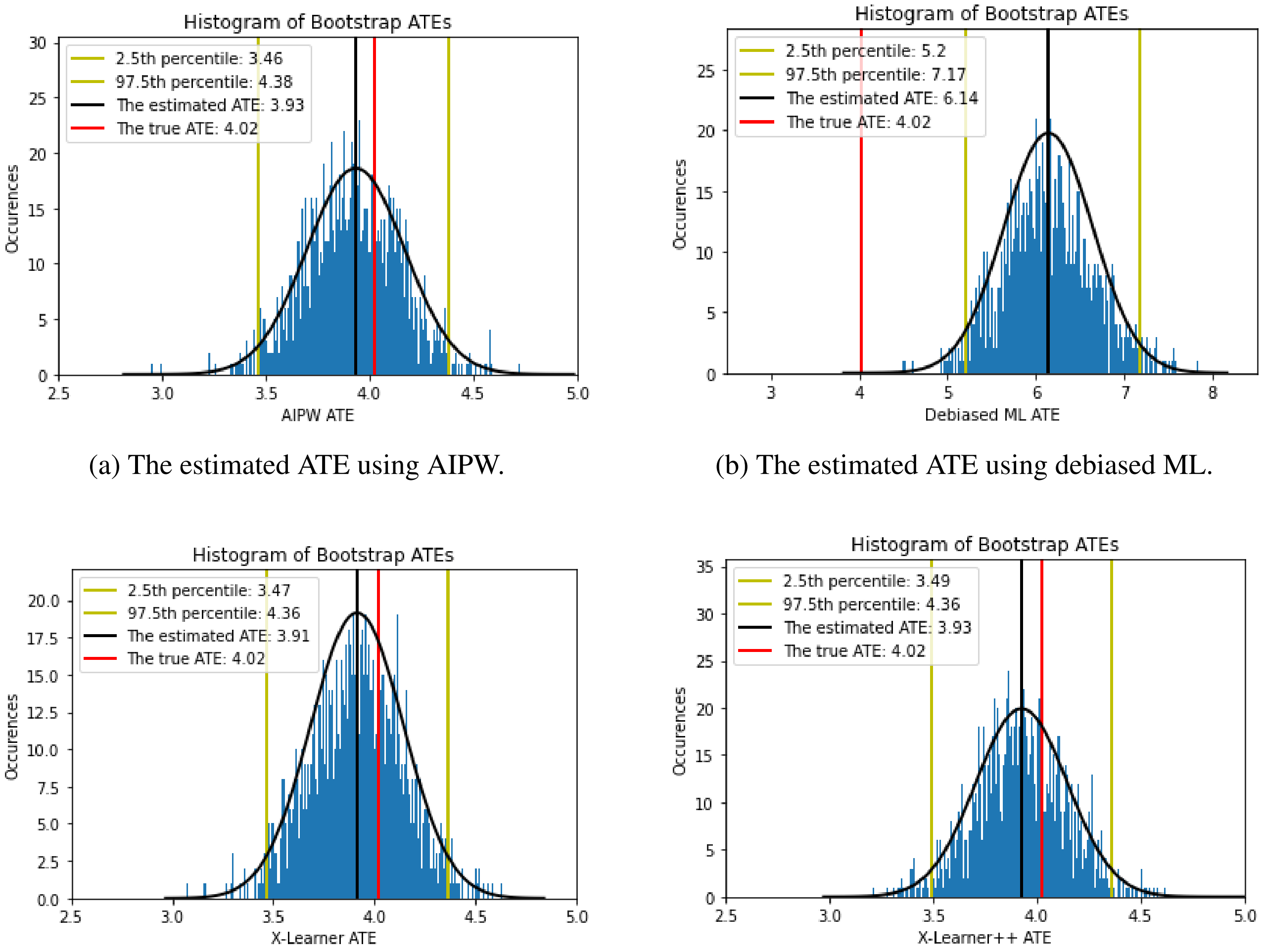
The results for the IHDP dataset are displayed in Table 9, and the corresponding histograms of the 5,000 bootstrap samples are displayed in Figure 5. The histograms also show the true ATE, the mean of the bootstrap samples, and the confidence intervals (CI). The estimated ATE using the AIPW estimator is 3.93, SD=0.24, CI=[3.46, 4.38]. The X-Learner estimator resulted in an ATE of 3.91, SD=0.23, CI=[3.47, 4.36], and the X-Learner++ yielded an ATE of 3.93, SD=0.22, CI=[3.49, 4.36]. The debiased ML estimator estimated an ATE of 6.14, SD=0.5, CI=[4.23, 5.1].

Estimator	ATE_{Est}	ATE_{GT}	ATE Error
AIPW	3.93	4.02	-0.09
X-Learner	3.91	4.02	-0.11
X-Learner++	3.93	4.02	-0.09
Debiased ML	6.14	4.02	2.12
Ensemble	4.66	4.02	0.64

Table 9: This Table shows the estimated ATE, and its ground truth, based on the synthetic dataset. The mean error for the ATE is presented.

5.3 Medical Dataset

The results for the VTE dataset are displayed in Table 10 and 11, for the ATE and CATE respectively. The histograms for the ATE are displayed in Figure 6. The ATE estimates are as follows: the AIPW estimator yielded an ATE of 3.13%, SD=4.34%, CI=[-5.34, 11.67], the debiased ML estimator resulted in an ATE of 4.23%, SD=5.92%, CI=[-7.45, 15.74], the X-Learner estimated an ATE of 0.17%, SD=4.32%, CI=[-8.42, 8.61], the extension X-Learner++ yielded 0.18% as ATE estimate, SD=4.3%, CI=[-8.25, 8.68], and the ensemble resulted in an ATE of 2.51%, SD=3.54%, and a CI of [-4.47, 9.45]. The CATE estimates are presented in Table 11.



(a) The estimated ATE using AIPW.

(b) The estimated ATE using debiased ML.

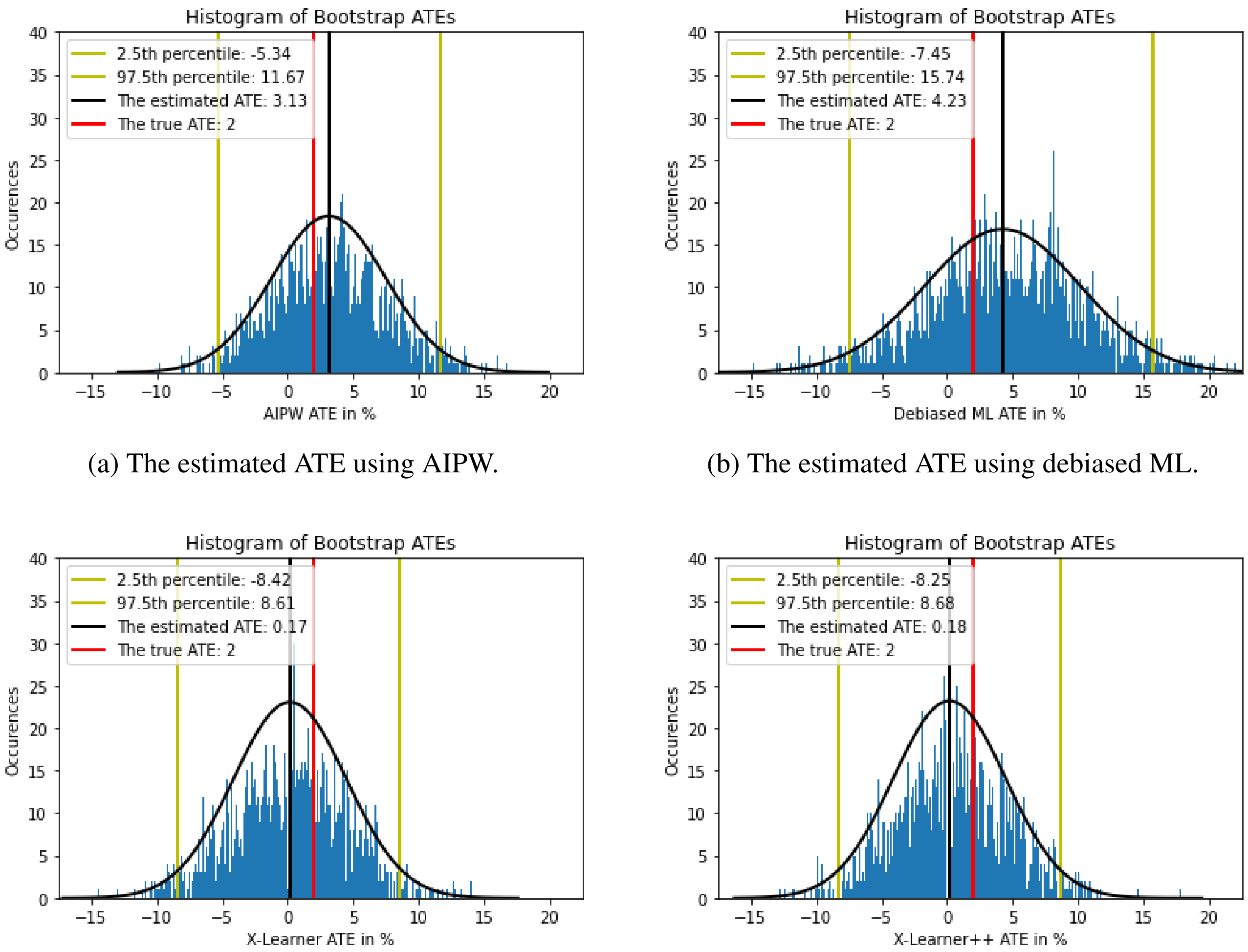
(c) The estimated ATE using X-Learner.

(d) The estimated ATE using X-Learner++.

Figure 5: The histograms of the bootstrap ATE distribution on the IHDP dataset. The red line corresponds to the true ATE, the black line to the mean of the 5,000 bootstrap samples, and the green lines correspond to the percentile confidence interval.

Estimator	ATE$_{Est}$	ATE$_{RCT}$	ATE Error
AIPW	3.13%	2%	1.13%
X-Learner	0.17%	2%	-1.83%
X-Learner++	0.18%	2%	-1.82%
Debiased ML	4.23%	2%	2.23%
Ensemble	2.51%	2%	0.51%

Table 10: This Table shows the estimated ATE and the ground truth ATE, which was derived from a comparable meta-analysis of multiple RCTs. Furthermore, calculated errors are displayed.



(a) The estimated ATE using AIPW.

(b) The estimated ATE using debiased ML.

(c) The estimated ATE using X-Learner.

(d) The estimated ATE using X-Learner++.

Figure 6: The histograms of the bootstrap ATE distribution on the VTE dataset. The red line corresponds to the ATE determined in RCTs, the black line to the mean of the 5,000 bootstrap samples, and the green lines correspond to the percentile confidence interval.

Estimator	CATE_{Est}	95% CI	CATE_{Est}(Statins=0)	95% CI
AIPW	3.08%	[-5.44, 11.5]	3.02%	[-3.32, 9.26]
X-Learner	-0.5%	[-12.65, 9.25]	-0.23%	[-5.64, 3.51]
X-Learner++	-0.5%	[-12.45, 8.66]	-0.31%	[-6.6, 3.15]
Debiased ML	3.93%	[-7.8, 15.75]	3.78%	[-4.54, 11.84]
Ensemble	2.17%	[-5.02, 8.94]	2.19%	[-2.44, 6.71]

Table 11: This Table shows the estimated CATE for the effect that Statin has on the efficacy of VTE chemoprophylaxis. There was no suitable ground truth accessible.

6 Discussion

6.1 IHDP

The estimation of the IHDP dataset shows the strength of causal reasoning, especially in situations where unconfoundedness is not violated. Despite the class imbalance, and the relatively low sample size, the AIPW estimator, the X-Learner, and the X-Learner++ yielded small errors when compared to the true ATE of this synthetic dataset. The low standard deviation of these estimators is indicative of the high certainty the estimators had in estimating the treatment effect, and the confidence intervals contain the true value. The debiased ML estimator showed the worst results, it overestimated the treatment effect and the true effect is outside of the bounds of the confidence intervals. This may be because the residuals of the treatment and outcome regressions have a non-linear relationship, and calculating the linear coefficient may not work very well. Consequently, the ensemble also overestimated the ATE, but less so due to the good performance of the other estimators included in the ensemble. While the ensemble does not yield the lowest error, it does protect against the misspecification of any single estimator in the ensemble.

The estimation on the IHDP dataset shows that the estimators are resilient toward a treatment selection bias, which is always apparent in observational data. This bias was induced by excluding a nonrandom portion of the data, as mentioned in Section 3. The strongest assumption underlying causal reasoning, unconfoundedness, can be discarded in this estimation because the original experiment was conducted as an RCT. Furthermore, it may be that the synthesization of the outcome to yield a ground truth may have simplified relations that would be much more complex in a real-world dataset. While the results are convincing, they must be interpreted with those limitations in mind that make the estimation of the ATE easier.

6.2 VTE

The results for the VTE dataset show much uncertainty in the estimation of the treatment effects, and no significant difference between the groups was found. A possible reason for this may be the small sample size in combination with the bootstrap approach. Some bootstrap samples may be overwhelmingly populated by patients who are especially at risk of VTE, or not as prone to VTE as the actual sample. Since the sub-sample in the ATE estimation that did develop VTE only contains 239 patients, the possible lack of heterogeneity in some bootstrap samples possibly leads to an over- or underestimation of the ATE. While the sample size in the IHDP dataset is balanced similarly, the synthesized outcomes probably ease the estimation. Furthermore, the included conditioning set for the VTE dataset only consists of four variables, two of which are binary. This may make the function approximation more difficult compared to the high-dimensional conditioning set in the IHDP dataset.

The CATE estimation conditional on prior use of Statins exacerbated the challenges already present in the ATE estimation. No significant difference between the two subgroups of Statin users and non-Statin users could be determined. Since the number of samples that suffered from VTE and consumed Statins prior is only 50, it may be challenging for a model to capture the trends within that subgroup. With the use of bootstrapping, the number of cases in some bootstrap samples is even lower than that. The comparison within the Statin consumers is therefore difficult.

There are two interconnected limitations that are noteworthy for this real-world dataset. The estimation relies on the completeness of the causal diagram presented in Figure 4. This is essential for the assumption of unconfoundedness, which is violated when a variable that influences the treatment selection and the outcome is omitted. A violation of this assumption leads to bias in the effect estimation. While the physicians at the SORG/FARIL research collaborative worked scrutinously on the presented diagram, a possible violation of the unconfoundedness assumption can not be ignored. Nevertheless, the replication of the effect estimated in the meta-analysis indicates that this may not be the case, or that the violation does not induce strong bias in the estimation.

While the high variance introduces uncertainty in the results, and no significant difference between the groups under investigation is found, a purely numerical comparison to a meta-analysis on the topic is interesting [51]. This meta-analysis found that both groups, the ones treated with VTE prophylaxis and the ones who were not, have a VTE incidence significantly different from zero. Furthermore, there is about a 2% difference (the VTE incidence is 12.5% for the treated, 10.5% for the untreated) between these two groups, albeit statistically insignificant. This 2% difference between the two groups is approximated in the results presented in Table 10, which show that the ensemble estimator benefits from overestimations in the AIPW and debiased ML estimators and an underestimation in the X-Learner and X-learner++. The difference between the two groups is insignificant in both, the meta-analysis and the results presented in this work. Therefore, this thesis replicated a part of the results of the meta-analysis, which concluded that VTE prophylaxis may not always be appropriate, especially after isolated ankle fractures due to the increase in the risk of adverse bleeding events following prophylactic treatment, and no significant effect of VTE prophylaxis. Nevertheless, the results presented here show high uncertainty and estimate the average treatment effect. Therefore, the evidence presented here requires further research, and patients at risk of VTE may still benefit from prophylaxis.

7 Conclusion

7.1 Summary of Main Contributions

This thesis contributes to the field of causal reasoning and the field of orthopaedics as follows. The development of an error correction for the X-Learner slightly reduces the error in the estimation of the treatment effect for the IHDP dataset, as compared to the X-Learner without the error correction. However, the error reduction is negligible, and, in the VTE dataset, the X-Learner++ performs slightly worse than the X-Learner. Generally, no significant difference between the X-Learner and the X-Learner++ could be found. Furthermore, the combination of multiple estimators to an ensemble can reduce the bias induced by some estimators overestimating the treatment effect, while others are underestimating the effect. The ensemble also reduces the variance in the estimation, and it performed best in the real-world dataset. Nevertheless, the bias in all estimators, including the ensemble, is high, which makes the results uncertain.

Another contribution of this thesis is the introduction of the causal reasoning methodology to the field of orthopaedics. A short version of this thesis is submitted to the Journal of Orthopaedic Research (JOR) to increase the exposure of physicians to this relatively new methodology. Furthermore, causal reasoning was introduced to the SORG/FARIL research collaborative at Harvard Medical School during the writing of this thesis through multiple presentations, the submitted research article, and the thesis itself. The evidence on the lack of efficacy of VTE prophylaxis in ankle fractures is strengthened and may influence clinical decision-making, albeit the high variance does not show confidence in the results. The thesis showed that causal reasoning can effectively replicate prior RCTs. The analysis of the effect of Statin consumption on VTE prophylaxis showed no significant differences between the groups. Therefore, there seems to be no effect of Statins on the efficacy of VTE prophylaxis. Nevertheless, we show that causal reasoning can emulate costly, timely, and sometimes unethical RCTs using widely available observational data.

7.2 Future Work

The field of causal reasoning can be supplemented by the complementary field of causal discovery. In causal discovery, the causal graph structure is determined in a data-driven way, not based on expert opinion. Advances in causal discovery may therefore make expert opinion obsolete, and reduce bias introduced by prior, personal conceptions of the world. This would enable us to decide on the structure of the causal diagram in a data-driven way without inducing the models with what we think causes what. While we attempted to utilize existing discovery algorithms, the results were not usable.

The assumption of unconfoundedness is strong and requires more attention. While the experiment on the IHDP dataset showed that the estimators can robustly determine the ATE, given suitable data and the absence of unobserved confounding. In contrast, the real-world dataset showed that the estimator's confidence can degrade quickly. It may be that there was unobserved confounding in the experiment, which can introduce bias. In other words, maybe there is another variable of which experts were not aware, which influences the treatment selection and the outcome. Furthermore, the small sample size, especially in the CATE analysis on Statin consumers, induces finite sample bias. The data is collected from one center, and bootstrapping with a small sample induced its own problems. A replication of the results with increased sample size and with multiple, geographically diverse centers may increase the generalizability and the certainty of the estimation.

Bibliography

[1] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. F. R. Jesus, R. F. Berriel, T. M. Paixão, F. W. Mutz, T. Oliveira-Santos, and A. F. de Souza, "Self-driving cars: A survey," *CoRR*, vol. abs/1901.04407, 2019.

[2] R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, Y. Zhou, C. Chang, I. Krivokon, W. Rusch, M. Pickett, K. S. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. H. Chi, and Q. Le, "Lamda: Language models for dialog applications," *CoRR*, vol. abs/2201.08239, 2022.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020.

[4] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *CoRR*, vol. abs/2102.12092, 2021.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

[6] R. Sutton, "The bitter lesson." http://www.incompleteideas.net/IncIdeas/BitterLesson.html, Mar. 2019. Accessed: 2022-11-30.

[7] M. Ford, *Architects of Intelligence*. Birmingham, UK: Packt Publishing, 2018.

[8] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics a primer*. Wiley, 2019.

[9] B. Schölkopf and J. von Kügelgen, "From Statistical to Causal Learning," Apr. 2022. arXiv:2204.00607 [cs, stat].

[10] D. B. Rubin, "[on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies," *Statistical Science*, vol. 5, pp. 472–480, Nov. 1990.

[11] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7353–7360, 2016.

[12] M. Hernán and J. Robins, *Causal Inference: What If*, vol. 1. Boca Raton: Chapman; Hall/CRC., 2020.

[13] M. A. Hernán and J. M. Robins, "Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available," *American Journal of Epidemiology*, vol. 183, pp. 758–764, Apr. 2016.

[14] M. A. Hernán and J. M. Robins, "Estimating causal effects from epidemiological data," *Journal of Epidemiology & Community Health*, vol. 60, pp. 578–586, July 2006. Publisher: BMJ Publishing Group Ltd Section: Continuing professional education.

[15] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, vol. 21, pp. C1–C68, 01 2018.

[16] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4156–4165, 2019.

[17] A. N. Glynn and K. M. Quinn, "An introduction to the augmented inverse propensity weighted estimator," *Political Analysis*, vol. 18, no. 1, p. 36–56, 2010.

[18] D. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.

[19] P. W. Holland, "Statistics and causal inference," *Journal of the American Statistical Association*, vol. 81, no. 396, p. 945–960, 1986.

[20] R. T. Gross, "Infant health and development program (IHDP): Enhancing the outcomes of low birth weight, premature infants in the united states, 1985-1988," *ICPSR Data Holdings*, 1993.

[21] R. Fischer, N. Nassour, B. Akhbari, C. W. DiGiovanni, J. H. Schwab, H. Ghaednia, and S. Ashkani-Esfahani, "Determining the causal effect of statins on reducing the incidence of venous thromboembolism after ankle fractures," *Journal of Orthopaedic Research*, submitted.

[22] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, (New York, NY, USA), p. 506–519, Association for Computing Machinery, 2017.

[23] F. H. Messerli, "Chocolate consumption, cognitive function, and nobel laureates," *New England Journal of Medicine*, vol. 367, no. 16, p. 1562–1564, 2012.

[24] H. Reichenbach and M. Reichenbach, *The Direction of Time*. Dover books on physics, Dover Publications, 1999.

[25] J. Pearl, "Bayesian networks: A model of self-activated memory for evidential reasoning," in *Proc. of Cognitive Science Society (CSS-7)*, 1985.

[26] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT press, 2nd ed., 2000.

[27] J. Pearl, *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd ed., 2009.

[28] J. Splawa-Neyman, "On the application of probability theory to agricultural experiments. essay on principles. section 9," *Statistical Science*, vol. 5, no. 4, pp. 465–472, 1990. Translated and edited by Dabrowska, D.M. and Speed, T.P.

[29] D. R. Cox, *Planning of Experiments*. New York: Wiley, 1958.

[30] J. Pearl, “The seven tools of causal inference, with reflections on machine learning,” *Communications of the ACM*, vol. 62, no. 3, pp. 54–60, 2019.

[31] J. Pearl and S. Russel, “Bayesian networks,” *Handbook of Brain Theory and Neural Networks, MIT Press*, p. 157–160, 2003.

[32] E. H. Simpson, “The interpretation of interaction in contingency tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 13, no. 2, p. 238–241, 1951.

[33] J. von Kuegelgen, L. Gresele, and B. Schoelkopf, “Simpson’s paradox in covid-19 case fatality rates: A mediation analysis of age-related causal effects,” *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 1, p. 18–27, 2021.

[34] J. Pearl, “Comment: Understanding Simpson’s paradox,” *The American Statistician*, vol. 68, no. 1, p. 8–13, 2014.

[35] M. A. Hernan, D. Clayton, and N. Keiding, “The Simpson’s paradox unraveled.,” *International Journal of Epidemiology*, vol. 40, p. 780–785, 2011.

[36] R. A. Fisher, “The design of experiments.,” *Oliver; Boyd, Edinburgh; London.*, vol. no. 2, 1937.

[37] E. Simpson, “The Interpretation of Interaction in Contingency Tables,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 13, no. 2, pp. 238–241, 1951.

[38] G. Imbens, “Nonparametric estimation of average treatment effects under exogeneity: A review,” *Review of Economics and Statistics*, 2004.

[39] J. K. Lunceford and M. Davidian, “Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study,” *Statistics in Medicine*, vol. 23, no. 19, pp. 2937–2960, 2004.

[40] G. King and L. Zeng, “The dangers of extreme counterfactuals,” *Political Analysis*, vol. 14, p. 131–159, 2006.

[41] M. C. Lovell, “A simple proof of the fwl theorem,” *The Journal of Economic Education*, vol. 39, no. 1, pp. 88–91, 2008.

[42] R. Frisch and F. V. Waugh, “Partial time regressions as compared with individual trends,” *Econometrica*, vol. 1, no. 4, pp. 387–401, 1933.

[43] G. U. Luwe, “On the theory of correlation for any number of variables, treated by a new system of notation,” *Proceedings of the Royal Society of London. Series A,*, vol. 79, no. 529, p. 182–193, 1907.

[44] J. L. Hill, “Bayesian nonparametric modeling for causal inference,” *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, p. 217–240, 2011.

[45] J. Brooks-Gunn, F. Liaw, and P. Klebanov, “Effects of early intervention on cognitive function of low birth weight preterm infants,,” *Pediatric Physical Therapy*, vol. 6, no. 1, 1994.

[46] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189 – 1232, 2001.

[47] J. Zhu, H. Zou, S. Rosset, and T. Hastie, “Multi-class Adaboost,” 2009.

[48] Office of the Surgeon General (US) and National Heart, Lung, and Blood Institute (US), *The Surgeon General's Call to Action to Prevent Deep Vein Thrombosis and Pulmonary Embolism*. Publications and Reports of the Surgeon General, Rockville (MD): Office of the Surgeon General (US), 2008.

[49] Z. Liederman, N. Chan, and V. Bhagirath, "Current Challenges in Diagnosis of Venous Thromboembolism," *Journal of Clinical Medicine*, vol. 9, p. 3509, Oct. 2020.

[50] E. Hawes and A. Viera, "Anticoagulation: Indications and risk classification schemes.," *FP essentials*, vol. 422, pp. 11–17, 07 2014.

[51] J. D. F. Calder, R. Freeman, E. Domeij-Arverud, C. N. van Dijk, and P. W. Ackermann, "Meta-analysis and suggested guidelines for prevention of venous thromboembolism (VTE) in foot and ankle surgery," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 24, pp. 1409–1420, 2016.

[52] A. Kotaska, "Venous thromboembolism prophylaxis may cause more harm than benefit: an evidence-based analysis of Canadian and international guidelines," *Thrombosis Journal*, vol. 16, p. 25, Oct. 2018.

[53] S. Gaertner, E.-M. Cordeanu, S. Nouri, C. Mirea, and D. Stephan, "Statins and prevention of venous thromboembolism: Myth or reality?," *Archives of Cardiovascular Diseases*, vol. 109, pp. 216–222, Mar. 2016.

[54] S. K. Kunutsor, S. Seidu, and K. Khunti, "Statins and primary prevention of venous thromboembolism: a systematic review and meta-analysis," *The Lancet. Haematology*, vol. 4, pp. e83–e93, Feb. 2017.

[55] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: what is it and how does it work?," *International Journal of Methods in Psychiatric Research*, vol. 20, pp. 40–49, Feb. 2011.

[56] P. N. Zivich and A. Breskin, "Machine learning for causal inference: On the use of cross-fit estimators," *Epidemiology*, vol. 32, no. 3, p. 393–401, 2021.

[57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

Appendices

A Submitted Paper

B Hyperparameters for GridSearch

B.1 Bagging

Parameters	Values
n_estimators	5, 25, 50, 100, 250, 500, 1000
warm_start	True, False

Table 12: The hyperparameters for the Random Forest base classifier and regressor used in the Bagging classifier and regressor.

B.2 Stacking

Parameters	Values
passthrough	True, False

Table 13: The hyperparameters for the Stacking classifier and regressor. The chosen stacking estimators were the Random Forest Classifier and regressor with 10 estimators, the Gradient Boosting classifier and regressor with a maximum depth of 10, a learning rate of 0.01, and 100 estimators, and the AdaBoost classifier and regressor with 100 estimators and a learning rate of 0.01.

B.3 MLP

Parameters	Values
hidden_layer_sizes	(2,3), (3,4,3), (8,4)
activation	tanh, relu
learning_rate_init	0.0001, 0.001, 0.01, 0.1, 0.4
learning_rate	constant, invscaling, adaptive
warm_start	True, False

Table 14: The hyperparameters for the Multi Layer Perceptron (MLP) classifier and regressor.

B.4 AdaBoost

Parameters	Values
n_estimators	5, 25, 50, 100, 250, 500, 1000
learning_rate	0.0001, 0.001, 0.01, 0.1, 0.4

Table 15: The hyperparameters for the AdaBoost classifier and regressor.

B.5 Random Forest

Parameters	Values
n_estimators	5, 25, 50, 100, 250, 500, 1000
max_depth	2, 5, 10, 25, 50
class_weight	balanced, balanced_subsample, none

Table 16: The hyperparameters for the Random Forest base classifier used in the Bagging classifier.

B.6 Gradient Boosting

Parameters	Values
n_estimators	5, 25, 50, 100, 250, 500, 1000
max_depth	2, 5, 10, 25, 50
learning_rate	0.0001, 0.001, 0.01, 0.1, 0.4

Table 17: The hyperparameters for the Gradient Boosting classifier.

B.7 Logistic Regression

Parameters	Values
solver	lbfgs, saga, elasticnet
penalty	l1, l2
class_weight	balanced, none

Table 18: The hyperparameters for the Logistic Regression classifier.

B.8 Support Vector Machine (SVM)

Parameters	Values
kernel	rbf, sigmoid, poly
gamma	auto, scale
C	0.1, 0.5, 1

Table 19: The hyperparameters for the SVM regressor.